THÈSE

Pour obtenir le grade de

Docteur

Délivré par

UNIVERSITE MONTPELLIER 2

Préparée au sein de l'école doctorale SIBAGHE

Et de l'unité de recherche BGPI

Spécialité : **Microbiologie/Parasitologie**

Présentée par **Pauline BERNARDO**

ÉCOLOGIE, DIVERSITÉ, ET DÉCOUVERTE DE PHYTOVIRUS À L'ÉCHELLE DE DEUX AGRO-ÉCOSYSTÈMES DANS UN CADRE SPATIO-TEMPOREL À L'AIDE DE LA GÉO-MÉTAGÉNOMIQUE

Soutenue le 19/12/2014 devant le jury composé de

Carolyn M. MALMSTROM, Pr. Michigan State University, USA	Rapporteur
Philippe BIAGINI, Dr. Directeur de recherche, CNRS, Marseille	Rapporteur
Thierry CANDRESSE, Dr. Directeur de recherche, INRA, Bordeaux	Examinateur
Claire NEEMA, Pr. Supagro, Montpellier	Examinateur
Catherine ENG, Ingénieur, Ministère de la défense, Paris	Invité
Philippe ROUMAGNAC, Dr. CIRAD, Montpellier	Directeur de thèse associé
Michel PETERSCHMITT, Dr. CIRAD, Montpellier	Directeur de thèse

Remerciements

Tout d'abord je tiens à remercier Carolyn Malmstrom, Philippe Biagini, Catherine Eng, Claire Neema et Thierry Candresse pour avoir accepté de faire partie de mon jury de thèse et pour le temps consacré à évaluer ce travail. Par ailleurs, je souhaite également remercier les membres de mon comité de thèse : Frédéric Fabre, Benoît Moury, Mohammed Barakat et Jean-Yves Rasplus pour m'avoir aiguillé dans les choix liés à ce projet.

Je remercie Philippe Rott ainsi que Claire Neema pour leur accueil au sein de l'UMR BGPI.

Merci à la DGA et à l'INRA d'avoir cru et accepté d'investir dans ce projet.

Je tiens notamment à remercier tous nos collaborateurs sans qui ce travail n'aurait pu aboutir. Mohammed Barakat et Philippe Ortet pour le traitement bioinformatique des séquences issues de NGS. Darren Martin pour son aide précieuse concernant les phylogénies, la recombinaison et l'évolution des géminivirus ainsi que pour son accueil à Cape Town. Gordon Harkins, conducteur hors-pair de 4x4, pour son aide lors des échantillonnages en Afrique du Sud. Arvind Varsani, le roi du clonage, pour son aide à l'obtention de génomes entiers de géminivirus. Nicole Yavercovski et Tony Rebello pour leurs identifications botaniques si précieuses. Damien Cohez pour son accueil et ses conseils à la Tour du Valat. Anna-Liisa Laine pour avoir partagé ses données de métagénomique avec nous et pour avoir eu la gentillesse de nous envoyer des échantillons de plantain finlandais.

Un merci incommensurable à Philippe Roumagnac, mon directeur de thèse, pour avoir encadré ce travail pendant plus de 3 ans, pour sa disponibilité, sa patience, son humanité, son humour, merci sincèrement pour m'avoir transmis son amour pour la recherche, sa motivation, et son enthousiasme sans pareil. Grâce à lui je pense être devenu quelqu'un d'autre et avoir grandi. Il a parfois mis ma patience à l'épreuve, j'ai parfois (souvent ?) mis la sienne à l'épreuve aussi, mais je pense qu'il sait désormais que j'aime relever les défis tout comme lui. J'ai souvent clamé que si ça n'avait pas été lui, je n'y serais pas arrivée… Mon avis est certainement subjectif (bien que je dispose de points de comparaison !), mais je pense sincèrement qu'il est le directeur de thèse, le directeur de stage et le chef d'équipe dont tout le monde rêve. Je pourrais le remercier encore longtemps pour tout ce qu'il m'a apporté, mais il faut en laisser pour les autres! Je tiens à remercier également Michel Peterschmitt, pour avoir co-encadré cette thèse, pour m'avoir guidé dans mes réflexions, pour ses idées et son grand enthousiasme. Merci à eux deux d'avoir fait de cette thèse ce qu'elle est, et merci pour le temps qu'ils m'ont consacré.

Merci aux membres de l'équipe 7 pour leur accueil, leur sympathie. Mille mercis Marie-Josée-Daroussat d'avoir été ma « mamie de substitution », merci pour son extrême gentillesse, son réconfort, toute son aide, ses gâteaux, sa paëlla, son humour… Toute thésarde rêverait d'avoir une collègue comme elle. Merci Denis Filloux pour ses conseils en bioinformatique ainsi que ses traitements de séquences sur GALAXY. Merci Jean Daugrois et Jean-Claude Girard pour leur sympathie. Merci Florence Barthod pour son accueil et toute son aide précieuse administrative. Merci Emmanuel Fernandez pour m'avoir initié aux techniques de biologie moléculaire et de m'avoir permis de voler de mes propres ailes au laboratoire, merci également pour sa sympathie et son humour. Merci Romain Ferdinand pour tout le matériel végétal qu'il nous a aidé à obtenir mais aussi pour sa gentillesse, sa bonne humeur, son humour et son enthousiasme.

Je remercie également mes deux stagiaires, Maëlle et Sarah, ce fut un énorme plaisir de les encadrer et de travailler avec elles.

Merci à tous les gens de l'UMR. Je tiens particulièrement à remercier Katia pour sa bonne humeur et ses conseils. Merci à mes garagistes BGPIens attitrés : Rémi, Loïc et Henri, sans vous je serais resté chez moi sans voiture ! Merci à Emmanuel Jacquot, Gaël Thébaud, François Bonnot, Sylvie Dallot et Nicolas Sauvion pour avoir été mon équipe de substitution durant les derniers mois de cette aventure et d'avoir fait de mes repas de midi une pause conviviale. Merci Martine Granier pour ton aide dans les manips. Merci Virginie Ravigné, une fois de plus Gaël Thébaud et Francois Bonnot pour leur aide précieuse et leurs conseils en statistiques. Merci encore à Emmanuel Jacquot pour avoir répondu à toutes mes questions concernant la thèse et m'avoir conseillé durant l'absence de Philippe.

Je remercie également Paul et Vellie pour leur accueil atypique mais plaisant dans la réserve de Bufflesfontein, ces souvenirs resteront gravés dans ma mémoire.

Charlotte (Mawie Théwése), merci pour sa complicité, son amitié, son humour, sa sympathie. Ces 3 années ont été si drôles à ses côtés ! Merci Loup, dieu d'Excel, des statistiques, et de la modélisation, ses conseils ont été précieux, mais là où je tiens à le remercier particulièrement c'est pour les moments de bêtises (et parfois de sérieux) que l'on a partagés. Emilie, c'est avec elle que la définition de la thèse en tant qu'aventure humaine prend tout son sens. Bien que tout nous sépare elle est pourtant devenue une de mes meilleures amies, ensemble nous avons ri, pleuré, nous nous sommes soutenues et conseillées. Je suis triste de savoir que je ne pourrai être présente pour la soutenir durant sa troisième année de thèse comme elle a pu le faire pour moi mais elle sait déjà que je penserai très fort à elle dans les moments difficiles comme dans les bons.

Un grand merci à mes amis : Marie je sais que même si tu es loin tu m'as toujours soutenue dans mes choix, Charly pour les moments de détente et de rire. Matthew, Dr Augé des Hautes Sphères, merci pour tous les services rendus pendant cette dernière année, pour son soutien moral, et un grand merci de m'avoir montré que je pouvais y arriver. Eline, cette dernière année à ses côtés a été magique, nous avons été là l'une pour l'autre que ce soit pour rire, pleurer mais encore et surtout danser ! J'ai également une pensée pour Benoît qui a supporté mes humeurs oscillantes au fil de la rédaction de ma thèse sans broncher, et merci pour m'avoir fait oublier par moments que j'étais en thèse, il a été mon voyage hors du temps.

Merci à mon frère Mickaël et ma belle-sœur Nadège pour avoir été présents dans les bons et les mauvais moments, leur présence m'a fait un bien énorme, et merci pour ce petit garçon Tony, mon neveu, ainsi que tout l'amour partagé. Merci à Lucas, mon adolescent de petit frère, pour sa fraîcheur, ses histoires et sa bêtise qui ont alimenté ma bonne humeur durant cette thèse.

Je tiens à m'excuser et à dire merci à tous ceux que j'aurais pu oublier.

Mon dernier remerciement mais aussi le plus grand et le plus chaleureux va à mes parents. Papa, Maman, merci pour tout l'amour dans lequel vous m'avez faite grandir. Sans vous je ne serais pas arrivée jusque-là. Merci de m'avoir soutenue et motivée. Merci d'avoir cru en moi et d'y croire encore. Je suis fière de vous et j'espère que vous serez encore plus fiers de moi. Papa, Maman, je vous dédie ce travail avec tout mon amour.

Résumé

Ecologie, diversité, et découverte de phytovirus à l'échelle de deux agro-écosystèmes dans un cadre spatio-temporel à l'aide de la géo-métagénomique

La connaissance de la diversité des phytovirus en milieu sauvage reste limitée. Les études concernant les interactions plantes-virus se sont en effet principalement focalisées sur les milieux cultivés. Ce manque de connaissance des milieux naturels et des interactions plante-virus qui s'y déroulent représente un écueil dans notre compréhension de l'écologie et de l'évolution des phytovirus sur le long terme. Cette quasi-absence de connaissance ne permet en outre pas de totalement comprendre, modéliser et prédire les processus micro- et/ou macro-évolutifs qui se mettent en place à l'échelle de l'agro-écosystème. Il est notamment encore difficile de quantifier l'impact des activités humaines (intensification de l'agriculture, transport de plantes, changement du climat, etc.) sur les interactions hôtes-agents pathogènes.

Une approche de géo-métagénomique a été développée dans deux agro-écosystèmes (le fynbos en Afrique du Sud et la Camargue) sur un pas de temps de deux ans. Cette approche nous a permis de ré-attribuer chaque séquence virale à son hôte géolocalisé. L'objectif de ce travail était d'évaluer (i) si le milieu sauvage constitue un réservoir de biodiversité phytovirale (ii) si il existe des patrons de distribution spatio-temporelle des phytovirus dans l'agro-écosystème et (ii) si des paramètres écologiques permettent d'expliquer ces distributions.

Grâce à cette nouvelle approche, une estimation de la diversité phytovirale associée aux deux agro-écosystèmes a pu être obtenue. Des patrons de distribution spatio-temporelle de plusieurs familles virales ont pu être mis en évidence. Les prévalences phytovirales associées au milieu cultivé se sont avérées être significativement plus importantes que celles associées au milieu non-cultivé pour trois des quatre campagnes d'échantillonnage. Par ailleurs, les plantes exotiques du fynbos sud-africain ont présenté des prévalences phytovirales significativement plus élevées que celles des plantes indigènes. Ces résultats soulignent l'impact direct ou indirect de l'activité humaine sur les dynamiques phytovirales à l'échelle de l'agro-écosystème.

Cette étude a également mené à la découverte potentielle de centaines de nouvelles espèces virales, dont trois nouvelles espèces appartenant à la famille des *Geminiviridae*. Ces espèces appartiennent à un nouveau genre au sein des *Geminiviridae* que nous avons nommé Capulavirus. Cette découverte nous a permis de mieux estimer certains paramètres liés à l'histoire évolutive des géminivirus (recombinaison, caractéristiques de leur ancêtre commun). Ce nouveau genre contient quatre espèces dont deux issues de plantes sauvages (*Euphorbia caput-medusae latent virus* et *Plantago Capulavirus*) et deux de plantes cultivées (*Alfafa leaf curl virus* et *French bean severe leaf curl virus*). Par ailleurs, nous avons obtenu des résultats préliminaires suggérant la transmission de ce nouveau genre par puceron, insecte qui n'a jamais été décrit comme vecteur de géminivirus. Ces découvertes nous ont amené à émettre des hypothèses sur l'émergence potentielle de ce nouveau genre à l'échelle mondiale.

Mots clés : Métagénomique, agro-écosystème, phytovirus, diversité, écologie, évolution, recombinaison, *Geminiviridae*, Capulavirus, plantes exotiques, plantes indigènes, fynbos, Afrique du Sud, Camargue.

Abstract

Ecology, diversity, and discovery of plant viruses over space and time across two agro-ecosystems using geo-metagenomics

Our knowledge about plant virus diversity in nature is still limited. Indeed, studies of plant-virus interactions have primarily focused on cultivated areas. This lack of knowledge about patterns of virus diversity and distribution in nature is hampering our understanding of plant virus ecology and evolution in the long term. In addition, this scarcity of knowledge does not allow to fully understand, model and predict the micro- and/or macro-evolutionary processes that are taking place across the agro-ecosystem. Consequently, it is still difficult to quantify the impact of human activities (agricultural intensification, plants transport, climate change, etc.) on host-pathogen interactions.

We developed a new metagenomics approach, the so-called geo-metagenomics approach, in order to provide information about the virus biodiversity, the prevalence of unknown and asymptomatic viruses and the spatial distributions of those plant viruses in two pilot ecosystems: the Western Cape Region of South Africa and the Camargue region in France. This approach provides geographically tagged cDNA from known and unknown viruses, and further allows linking viral sequences obtained by the metagenomics approach to a specific host, and hence to geographic coordinates. The objectives of this study were to assess (i) if wild areas can be considered as reservoir of plant virus biodiversity (ii) if there exists patterns of spatio-temporal distribution of plant viruses at the agro-ecosystem scale and (ii) if ecological parameters can account for these distributions.

This new approach allowed us to estimate plant virus diversity associated with two agro-ecosystems. Patterns of spatial and temporal distribution of several viral families have been highlighted. Plant virus prevalences associated with cultivated areas were found to be significantly greater than those associated with non-cultivated areas for three out of the four sampling surveys. Furthermore, exotic plants from South African fynbos showed significantly higher prevalence than native plants. These results emphasize the direct or indirect impact of human activity on plant virus dynamics at the agro-ecosystem scale.

This study also led to the discovery of hundreds of potential new viral species, including three new species belonging to the family *Geminiviridae*. These species belong to a new genus of the *Geminiviridae* family that we tentatively named Capulavirus. This discovery sheds a new light on the evolutionary history of geminiviruses (recombination, genomic features of their common ancestor). This new genus contains four species, including two species isolated from wild plants (*Euphorbia caput-medusae latent virus* and *Plantago Capulavirus*) and two species recovered from crops (*Alfalfa leaf curl virus* and *French bean severe leaf curl virus*). Finally, our results suggest that aphids may transmit virus species from this new genus, which has never been described so far for geminiviruses vection. The potential emergence of this new genus is finally discussed.

Keywords: Metagenomics, agro-ecosystem, plant viruses, diversity, ecology, evolution, recombination, *Geminiviridae*, Capulavirus, exotic/native plants, fynbos, Camargue.

Table des matières

Liste des figures et tableaux

Figures de la Synthèse bibliographique

Tableaux de la Synthèse bibliographique

Figures du Chapitre I

Tableaux du Chapitre I

Figures du Chapitre II

Tableaux du Chapitre II

Liste des abréviations

454 : Pyroséquençage Roche 454
ACP : Analyse en composante principale
ADN : Acide Désoxyribonucléique
AFD : Analyse Factorielle Discriminante
ALCV : *Alfafa leaf curl virus*
ARN: Acide Ribonucléique
***avr* :** gène d'avirulence
BYDV : *Barley yellow dwarf virus* (BYDV-PAV est une des souches de BYDV)
C2010/C2012 : Echantillonnage Camargue en 2010/2012
C2012 : Echantillonnage Camargue 2012
***cp*/CP :** gène codant pour la protéine de capside/protéine de capside
CYDV: *Cereal yellow dwarf virus*
dsDNA: ADN double brin
EcmLV: *Euphorbia caput-medusae latent virus*
F2010/F2012 : Echantillonnage fynbos en 2010/2012
FbSLCV: *French bean severe leaf curl virus*
ICTV: International Committee on Taxonomy of Viruses
LIR: Large Intergenic Region
mARNs: ARN messagers
MID: Molecular Identifier
miRNAs: micro ARNs
MSV:*Maïze streak virus*
NCBI : National Center for Biotechnology Information
NGS : Next Generation Sequencing
ORF : Open Reading Frame (Cadre de lecture ouvert)
OTU : Operational Taxonomic Unit
pb : paire de bases
PCR : Polymerase Chain Reaction
RCA : Rolling Circle Amplification
***rep*/Rep :** gène codant pour la protéine associée à la réplication/protéine associée à la réplication
res : gène de résistance
rPCR : Random Polymerase Chain Reaction
rt-PCR : Reverse Transcription Polymerase Chain Reaction
SIR : Short Intergenic Region
siRNAs : petits ARNs interférents
ssDNA : ADN simple brin
ssRNA : ARN simple brin
ssRNA- : ARN simple brin à polarité négative
ssRNA+ : ARN simple brin à polarité positive

Synthèse Bibliographique

1. Diversité des virus

1.1. Qu'est ce qu'un virus ?

1.1.1. Les virus : des entités biologiques

1.1.1.1. Bref historique de la virologie des plantes

Même si au VIII^{ème} siècle des écrits décrivent des symptômes sur *Eupatorium sp.* causés par l'agent que l'on connaît aujourd'hui comme étant *Eupatorium yellow vein virus* (Saunders *et al.*, 2003), ce n'est qu'à la fin du XIX^{ème} siècle que nait la virologie. En 1892, le russe Dimitri Ivanovsky démontre que des extraits de tabacs atteints de la mosaïque du tabac filtrés à travers un filtre de Chamberland (filtre qui ne laisse pas passer les bactéries) peuvent transmettre la maladie à des plants de tabac sains. En 1898, Martinus Willem Beijerinck observe en outre que cet agent peut diffuser à travers l'agar (qui retient les bactéries) et ne réussit à le multiplier que dans des plantes vivantes. Il donne alors le nom de virus à cet agent infectieux (qui provient du latin « *virus* » qui signifie « poison, toxine ») et le définit comme «*contagium vivum fluidum*» (fluide vivant contagieux) (Creager *et al.*, 1999). Martinus Willem Beijerinck introduit ainsi l'idée qu'un virus est virulent et de petite taille.

En 1935, Wendell Stanley décrit la cristallisation du *Tobacco mosaïc virus* (TMV). Ces cristaux sont à 90% constitués de protéines, 0.5% de phosphore et 5% d'ARN ; ces résultats posaient la question fondamentale du mécanisme de multiplication de ces nouvelles entités. Wendell Stanley définit alors le TMV comme une « protéine autocatalytique » qui nécessite la présence de cellules vivantes pour se multiplier (Stanley, 1935). Ces résultats pionniers ont ouvert la voie à la virologie moderne et ont initié une question qui est toujours d'actualité: les virus sont-ils des organismes vivants et quelle est leur identité ?

1.1.1.2. Débat sur l'identité virale

Depuis une centaine d'années, la communauté scientifique n'a cessé de débattre sémantiquement de l'identité virale (Villarreal, 2008). Un premier terme qui apparaît souvent dans la définition des virus est celui d'«entité» ; ici nous définirons l'entité en tant que « chose » ayant son individualité. De plus, les virus ont souvent été définis comme entité biologique, ce qui signifie que ce sont des êtres qui relèvent du domaine du vivant, or le fait que les virus soient vivants ou non est à l'origine de nombreux débats. Il est ainsi important de définir ce qu'est un être vivant. Depuis Aristote (350 avant JC) jusqu'à nos jours cette définition n'a cessé d'évoluer mais une caractéristique semble retenue quelle que soit la définition : un être vivant a la capacité de se répliquer, c'est-à-dire de produire de nouveaux individus (Moreira and Lopez-Garcia, 2009). On peut également ajouter à la vie des caractéristiques indiscutables qui sont ses limitations

à la naissance et à la mort. De plus, les organismes vivants ont un certain degré d'autonomie biochimique permettant des activités métaboliques amenant à la survie de l'organisme. Cependant, en ce qui concerne ces activités métaboliques, les virus sont dépendants de l'hôte. L'aspect « vivant » du virus dépend alors de son potentiel à se retrouver dans un hôte.

Des travaux initiés par Wendell Stanley et d'autres chercheurs ont démontré qu'un virus, sous la forme d'un virion, est constitué d'acide nucléique (ADN ou ARN) renfermé dans une ou plusieurs protéines de capside pouvant parfois abriter des protéines impliquées dans l'infection. Cette définition des virus les classe ainsi dans le monde chimique plutôt que biologique. Pourtant, si on se focalise sur l'activité d'un virus dans une cellule ce dernier est loin d'être inerte. En effet, une fois dans la cellule, un virus perd sa capside, libère son génome et va induire sa réplication et la synthèse de ses protéines en utilisant la machinerie cellulaire hôte. Les virus sont aussi considérés comme étant des entités biologiques dans le sens où ils possèdent un génome et sont capables de s'adapter à leur hôte et à leur environnement. Nous considèrerons ainsi les virus comme des entités à la frontière entre le vivant et l'inerte : ils ne peuvent pas se multiplier par eux-mêmes mais peuvent le faire dans une cellule vivante. Les virus peuvent profondément affecter leur hôte jouant ainsi un rôle prépondérant dans l'histoire évolutive des organismes. La découverte du virophage Sputnik (virus parasite de Mimivirus) a montré qu'un virus peut être parasité par un autre virus, suggérant que les virus devraient être classés dans le domaine du vivant (Pearson, 2008). La définition de l'identité virale dépend ainsi du cadre dans lequel ils sont observés : infection dans la cellule ou juste en tant qu'entité. Marc Van Regenmortel et Brian Mahy ont défini récemment les virus comme des « formes de vie empruntées » et Ed Rybicki les a qualifiés d'« organisms at the edge of life » (Rybicki, 1990; Van Regenmortel, 2000). Patrick Forterre est quant à lui plus radical, il considère que le monde vivant peut être divisé en deux groupes : d'un côté, les organismes cellulaires et de l'autre, les virus (Forterre, 2010). L'ensemble de ces travaux et opinions montrent que la notion du vivant est une notion dynamique, évoluant en fonction de nos connaissances (Saïb, 2006). De façon paradoxale, les travaux de virologie, et les définitions fluctuantes des « virus » devraient permettre d'éclairer les débats sur le « vivant » et sur l'origine de la vie.

1.1.2. Les virus : une identité biologique spécifique

Trois critères permettent de distinguer les virus des organismes cellulaires eucaryotes et procaryotes (Hull, 2002):

(1) Il n'y a pas de membrane qui sépare le virus de la cellule hôte durant la réplication,

(2) Les virus ne possèdent pas de système complet de synthèse de protéines,

(3) La réplication des virus se fait via la synthèse d'un ensemble (pool) de composants suivie par l'assemblage de particules virales à partir de cet ensemble (Figure SB.1b).

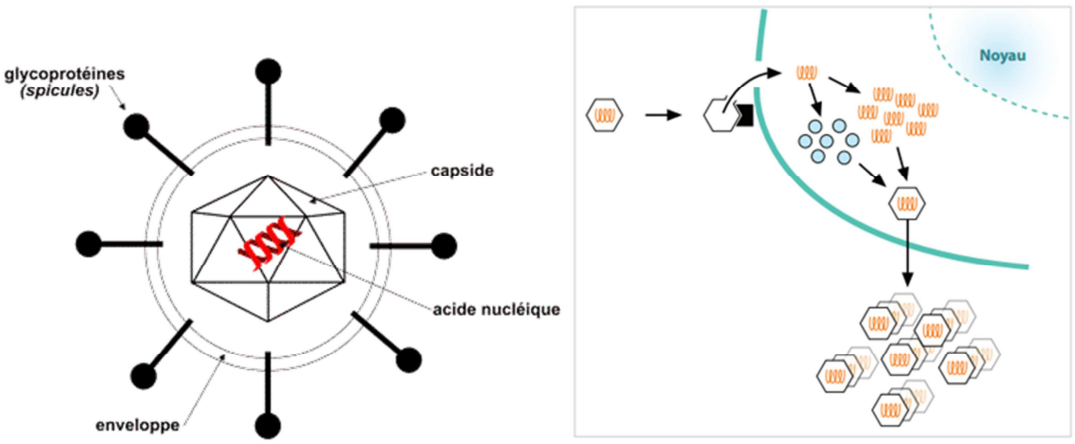


Figure SB.1 :a)Schématisation simplifiée de la structure d'un virus. b) Schématisation simplifiée de l'infection d'une cellule par un virus. Lors de l'entrée du virus dans la cellule, son génome est décapsidé et répliqué, les protéines de la capside sont également synthétisées à partir de ce génome et finalement ce pool d'acides nucléiques et de protéines virales va servir à créer de nouvelles particules virales qui vont être libérées de la cellule hôte pour en infecter d'autres. Sources : http://www.museum-grenoble.fr/passe/sciencefete/3/structure.html; http://www.afd-ld.org/~fdp_viro/pdf/chap1.pdf.

À partir de ces critères et de la discussion précédente sur l'identité virale nous proposons de donner la définition suivante d'un virus : un virus est une entité acellulaire ayant un génome composé d'une ou plusieurs molécules d'acide nucléique (ADN ou ARN) qui code au moins pour une protéine impliquée dans sa réplication (ce qui le différencie des plasmides, transposons et viroïdes) et qui, une fois dans la cellule hôte, peut induire sa propre multiplication. La réplication virale est donc dépendante de la machinerie cellulaire de l'hôte, elle a lieu dans le cytoplasme ou dans le noyau de la cellule hôte, de façon libre sans séparation via une quelconque membrane (Figure SB.1b). Une fois synthétisé, le génome viral est recouvert d'une coque protéique appelée capside qui est un assemblage d'éléments généralement organisés en formes géométriques régulières (hélice ou icosaèdre) et par-dessus, dans certains cas, d'une enveloppe externe provenant des membranes de l'hôte (Figure SB.1a).

Les génomes viraux arborent une grande diversité de nature et de structure (Figure SB.2), en effet, il existe deux formes de support de l'information génétique (ADN et ARN) déclinés en plusieurs variantes: ARN simple brin positif, ARN simple brin négatif, ARN double brin, ADN simple brin, ADN double brin rétroïde, et ADN double brin non-rétroïde (ce dernier n'a jamais été observé chez les virus de plantes). Le génome des virus peut comporter de 1 (virus Delta, parasite du virus de l'Hépatite B) à 2500 gènes (Pandoravirus). La taille des particules virales peut aller de 18 nm de diamètre (Nanovirus) à une largeur de 0,5 micromètre et une longueur de 1,5µm (Pithovirus).

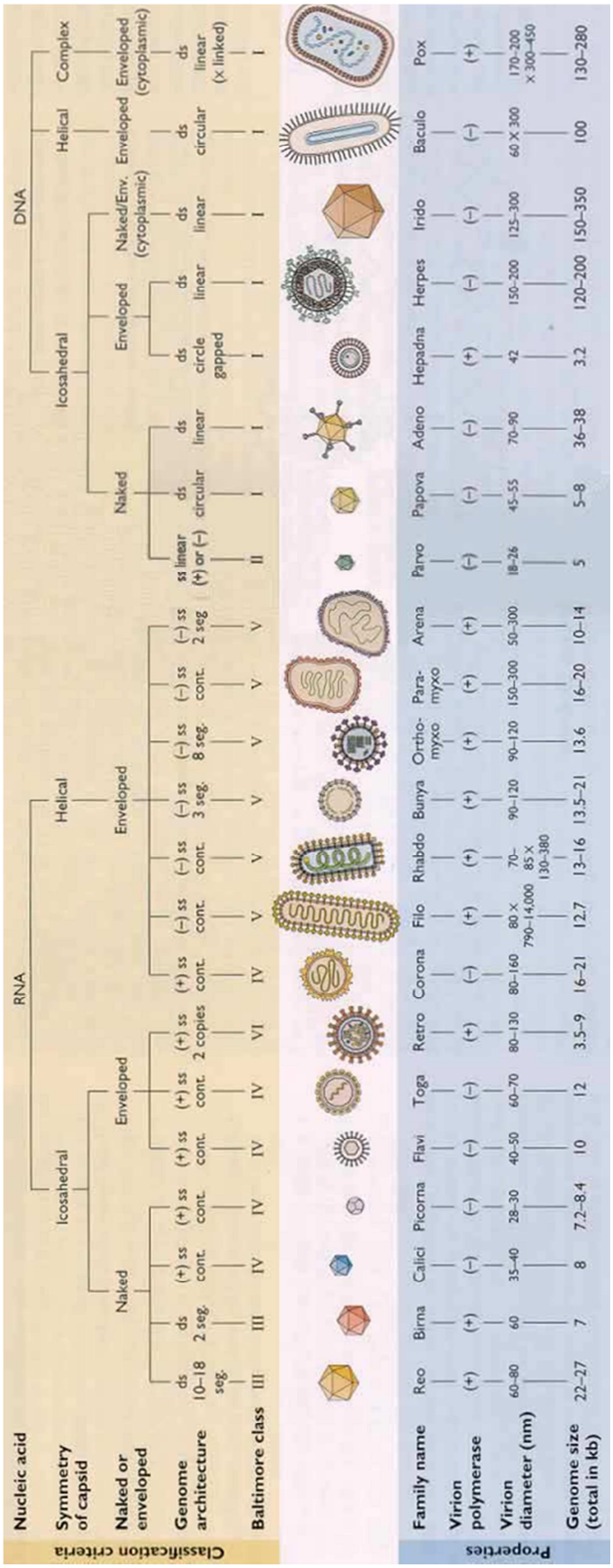


Figure SB.2 : Représentation non exhaustive des différentes formes de virus. Source : ICTV.

1.2. Estimation de la diversité des virus

Il est aujourd'hui considéré que les virus sont les entités biologiques les plus abondantes et les plus diversifiées sur terre. Plusieurs articles citent en référence un article de Mya Breitbart et Forest Rohwer de 2005 qui avance qu'il y aurait « 10^{31} virus sur Terre » (Breitbart and Rohwer, 2005). Toutefois, ce chiffre, bien que devenu très populaire, reste une approximation très discutable de travaux de comptage de particules assimilées à des virus (virus-like particles) réalisés grâce à de la microscopie à épifluorescence à partir d'échantillons d'eau de mer (Breitbart and Rohwer, 2005). Par ailleurs, il est difficile de savoir si ce chiffre évoque un nombre d'espèces ou une quantité de particules virales. Il semble donc encore difficile d'avoir une vision précise de la diversité globale des virus. Les virus sont présents dans une large gamme d'organismes (bactéries, archées, plantes, champignons, animaux) occupant des niches écologiques très variées (océans, sol, aérosol etc.) (Adriaenssens *et al.*, 2014; Blinkova *et al.*, 2010; Donaldson *et al.*, 2010; Hall *et al.*, 2013; Kim *et al.*, 2008; Ng *et al.*, 2011b; Roossinck *et al.*, 2010; Rosario *et al.*, 2009; Roux, 2012; Singh *et al.*, 2012; van den Brand *et al.*, 2012; Victoria *et al.*, 2009; Wong *et al.*, 2012). Toutefois, les travaux de métagénomique ont démontré que les communautés virales environnementales (issues de sols, eaux, aérosols, etc.) sont différentes des virus précédemment caractérisés (à partir d'individus eucaryotes ou procaryotes). Ces études reportent en outre plus de 50% de séquences considérées comme inconnues (sur la base de leur similarité avec les séquences déposées dans la banque internationale de donnée GenBank). Par ailleurs, de nombreuses séquences « environnementales » présentent des similarités faibles avec des séquences de la GenBank, ce qui laisse supposer que ces fragments nucléiques appartiennent à de nouvelles espèces virales non encore décrites (Rosario and Breitbart, 2011). Ces résultats ont suggéré que la diversité virale est significativement plus importante que la diversité bactérienne (Rosario and Breitbart, 2011).

1.3. Moteurs moléculaires de la diversité virale

Les génomes viraux varient au cours du temps. Les nouveaux génotypes ainsi générés sont soumis à la sélection naturelle. Cette variabilité génétique permet l'adaptation aux hôtes, notamment par contournement de résistances et évolution de la virulence (Acosta-Leal *et al.*, 2011). Même si dans un même hôte on peut constater la présence de plusieurs variants génétiques (génotypes viraux), il est aujourd'hui reconnu que la diversité génétique des populations virales inter-hôte est beaucoup plus élevée que la diversité génétique intra-hôte (Garcia-Arenal *et al.*, 2001). Il existe trois moteurs moléculaires de la diversité virale : la mutation, la recombinaison et le réassortiment.

1.3.1. La mutation

La mutation est la résultante d'une copie imparfaite du matériel génétique des parents à la descendance ou de phénomènes physiques ou chimiques (mutations par contact avec les UV par exemple). Ces phénomènes induisent des modifications dans la chimie des bases nucléotidiques (Acosta-Leal *et al.*, 2011). La mutation peut ainsi

survenir lorsque la polymérase fait des erreurs au cours de la réplication des génomes. Ce sont des erreurs ponctuelles qui existent sous trois formes :

(1)L'insertion qui correspond à l'insertion d'un ou plusieurs nucléotides entre deux nucléotides préexistants,

(2)La délétion qui correspond à la perte d'un ou plusieurs nucléotides dans le génome viral,

(3)La substitution qui correspond au remplacement d'un nucléotide par un autre. Cette forme de mutation est quatre fois plus fréquente que les deux décrites précédemment.

Le taux de mutation indique les fréquences de modifications générées par la réplication. Pour illustrer les vitesses auxquelles les virus évoluent on utilise fréquemment le taux de substitution. Le taux de substitution est défini ici comme la probabilité de l'occurrence d'une substitution nucléotidique par nucléotide par cellule infectée (s/n/c) (Sanjuan *et al.*, 2010). Il est plus élevé chez les virus que chez les organismes cellulaires, et il semblerait y avoir une corrélation négative entre le taux de mutation et la taille du génome (Figure SB.3) (Drake, 1991; Sanjuan *et al.*, 2010). Les taux de mutation vont de 10^{-8} à 10^{-6} s/n/c pour les virus à ADN et de 10^{-6} à 10^{-4}s/n/c pour les virus à ARN (Sanjuan *et al.*, 2010). Par ailleurs, la fréquence de mutation des phytovirus dépend fortement de l'espèce hôte (Schneider and Roossinck, 2001). Les individus ayant un fort taux de mutation sont appelés « mutateurs » et ceux avec un faible taux de mutation sont appelés « antimutateurs » (Mansky and Cunningham, 2000), ces caractéristiques vont jouer un rôle clé dans l'évolution, la virulence et l'émergence virale.

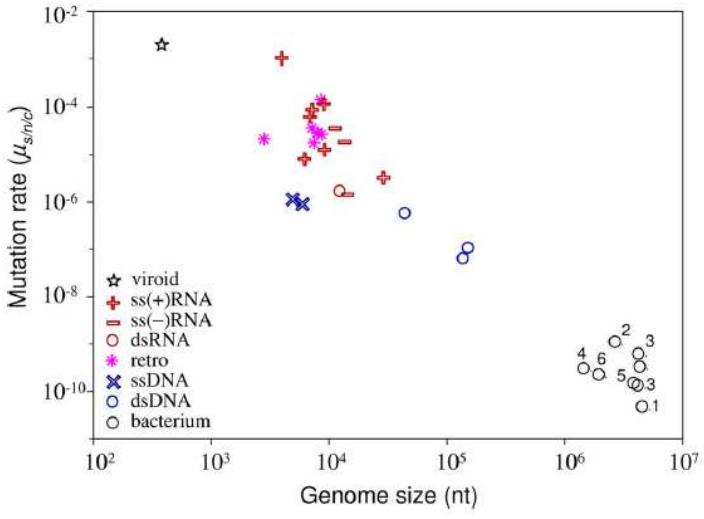


Figure SB.3 : Relation entre le taux de mutation et la taille du génome. Les groupes majeurs de virus sont indiqués. Les valeurs pour les bactéries et les viroïdes sont également indiquées à titre de comparaison. D'après Sanjuan *et al.*, 2010

Les mutations peuvent avoir trois types d'effet en termes évolutifs : délétère, neutre et bénéfique. Comme leur nom l'indique les mutations bénéfiques vont apporter un avantage sélectif contrairement aux mutations délétères qui sont désavantageuses.

La mutation neutre quant à elle n'apporte aucun avantage ni désavantage. La majorité des mutations (70%) sont délétères voire létales, par conséquent, tous les variants produits dans une population virale ne sont pas maintenus (Domingo-Calap *et al.*, 2009; Sanjuan *et al.*, 2004). Cependant, certaines variations délétères dans un environnement donné peuvent être bénéfiques (adaptatives) dans un autre (Duffy *et al.*, 2006; Ferris *et al.*, 2007). Ce phénomène, appelé pléiotropie antagoniste, se traduit par une corrélation génétique négative des performances des individus dans deux niches différentes (Levins, 1968).

L'analyse des mutations va pouvoir nous informer sur la sélection qu'elles subissent. Une des mesures permettant cette analyse est le ratio dN/dS, où dN correspond au taux de substitution des mutations non-synonymes (mutations conduisant à la synthèse d'un acide aminé différent) et dS au taux de substitution des mutations synonymes (mutations conduisant à la formation du même acide aminé). Ce ratio donne ainsi une indication du type de sélection subie par les séquences étudiées. Classiquement on considère que les séquences étudiées sont soumises à une sélection positive/adaptative lorsque ce $dN/dS >> 1$, à une sélection négative lorsque $dN/dS << 1$, et à une sélection neutre lorsque $dN/dS = 1$. A titre d'exemple, une population de phytovirus avec un ratio de 0.01 à 0.31 sera considérée comme une population en équilibre adaptée à son hôte (Garcia-Arenal *et al.*, 2001).

1.3.2. La recombinaison

Lorsque deux virus infectent une même cellule, il peut y avoir échange de gènes ou plus généralement de parties de gènes entre ces deux virus. La recombinaison est la formation de molécules d'acides nucléiques chimériques filles à partir de molécules d'acides nucléiques provenant de parents différents. La molécule fille correspond donc à une mise en continuité du matériel génétique provenant de deux chaines différentes d'acide nucléique. C'est un mécanisme qui survient durant la réplication. Des phénomènes de recombinaison ont été détectés entre espèces, genres et même récemment entre familles différentes (Saunders and Stanley, 1999). Il existe deux types de recombinaison :

- La recombinaison homologue : le changement de chaîne se produit au niveau de sites homologues (Figure SB.4a),

- La recombinaison non-homologue : la recombinaison se produit entre des sites non homologues (Figure SB.4b).

La fréquence de recombinaison par nucléotide est de l'ordre de 10^{-4} à 10^{-8} et toutes les régions d'un génome ne sont pas soumises de la même manière à ce phénomène ; il existe des régions où le taux de recombinaison est élevé ('hotspots') et d'autres où le taux de recombinaison est faible ('coldspots') (Garcia-Arenal *et al.*, 2001). Par exemple, il a été montré que des génomes de begomovirus ont une distribution de

points de recombinaison non aléatoire qui est soumise à la sélection (Lefeuvre *et al.*, 2007a; Martin *et al.*, 2011b).

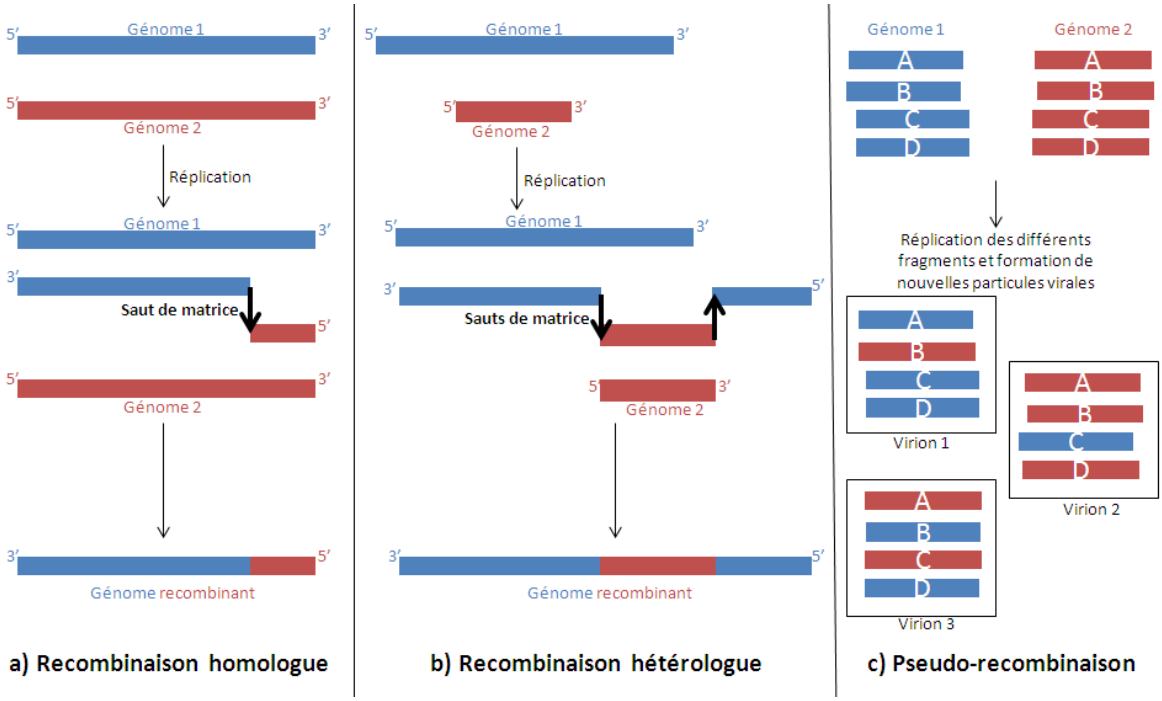


Figure SB.4: Les différents types de recombinaison. a) La recombinaison homologue : lors de la réplication la polymérase va effectuer un changement de matrice entre deux brins pratiquement identiques. **b) La recombinaison hétérologue** : lors de la réplication la polymérase va effectuer un changement de matrice entredeux brins différents. **c) La pseudo-recombinaison** : lors de la co-infection d'une cellule par deux virus ayant des génomes segmentés différents, les deux génomes vont être répliqués, et des fragments issus des deux génomes peuvent être encapsidés dans la même particule virale (par exemple le virion 1 est constitué des fragments A, C et D issus du génome 1 et du fragment B issu du génome 2).

La recombinaison est une source importante de variation génomique. De ce fait, elle peut avoir un impact important sur l'évolution des virus ; elle est à l'origine de nouvelles souches, espèces et même de nouveaux genres (Garcia-Andres *et al.*, 2006; Martin *et al.*, 2011a; Padidam *et al.*, 1999). Dans la majorité des cas, les phénomènes de recombinaison se traduisent par la formation de virus défectifs à non viables (Acosta-Leal *et al.*, 2011). Toutefois, une étude focalisée sur des begomovirus recombinants a montré que ces virus avaient des niveaux de virulence équivalents ou intermédiaires à leurs deux parents (Vuillaume *et al.*, 2011). Par ailleurs, le fait que la recombinaison puisse amener à la formation de nouveaux genres (par exemple, chez les *Geminiviridae*, les becurtovirus sont issus de la recombinaison entre begomovirus et curtovirus) rend difficile l'établissement de phylogénies.

1.3.3. Le réassortiment

Lors de l'infection d'un hôte par des virus phylogénétiquement proches dont les génomes sont composés de plusieurs molécules d'acide nucléique, il arrive que ces virus échangent certaines de ces molécules ; on appelle cela le réassortiment ou la pseudo-recombinaison (Figure SB.4c). Tout comme la recombinaison, ce phénomène permet

l'apport de nouveaux gènes, ce qui aura indéniablement un fort impact potentiel sur l'évolution des virus. Chez les phytovirus, les nanovirus, qui ont des génomes segmentés, sont connus pour être soumis au réassortiment des différents fragments génomiques qui les composent (Grigoras *et al.*, 2014).

1.4. Taxonomie des virus

La taxonomie des virus a pour but de nommer les entités virales de manière pertinente et universelle et de les classer dans des groupes de la façon la plus rationnelle possible (Astier, 2007). La classification des virus est organisée par le Comité International de Taxonomie des Virus (ICTV). Elle est basée sur la classification de Baltimore (biologiste américain) et s'organise de manière hiérarchique. Les virus sont tout d'abord séparés en fonction de la nature et de la structure de leur acide nucléique ainsi que de leur stratégie de réplication puis ils sont classés en 4 taxa : ordre, famille, genre et espèce.

1.4.1. L'espèce

La définition de l'espèce virale adoptée par l'ICTV en 1991 est celle de Marc Van Regenmortel (1989) : « L'espèce virale est un ensemble polythétique de virus qui constitue une lignée de réplication et occupe une niche écologique particulière. » (Tableau SB.1) (Van Regenmortel, 1989). Une espèce virale est donc considérée comme étant une entité biologique et écologique. Les caractéristiques permettant de délimiter les espèces virales sont : l'organisation du génome (caractéristique majeure), la structure et la physico-chimie du virion, les propriétés sérologiques de la capside, la gamme d'hôte, la vection, et les relations avec l'hôte (Astier, 2007). Chaque espèce a ses caractéristiques propres qui la définissent et qui sont déterminées par des virologues spécialistes (Van Regenmortel *et al.*, 1997). Pour appartenir à une espèce, le virus n'est pas obligé de remplir chacun des critères cités précédemment, c'est le caractère polythétique de l'espèce (Tableau SB.1).

Toutefois, la définition de l'espèce virale établie par l'ICTV a récemment fait l'objet d'une polémique (Gibbs and Gibbs, 2006; Van Regenmortel *et al.*, 2013) concernant le maintien ou la suppression du terme « polythétique » (Tableau SB.1). Cette polémique est née lorsque plusieurs groupes au sein de l'ICTV, notamment les groupes travaillant sur les *Geminiviridae,* ont décidé d'établir les nouvelles propositions d'espèces sur la base exclusive de seuils d'identité de séquences, i.e. un caractère monothétique. Ces sous-groupe de l'ICTV ont justifié ce choix par le fait que les comparaisons de séquences sont objectives alors que les informations découlant d'observations visuelles telles que la symptomatologie sont subjectives (Varsani *et al.*, 2014a).

Le dernier rapport de l'ICTV (2013) répertorie un total de 2827 espèces virales. Par ailleurs, il est ici important de souligner qu'un isolat désigne généralement un échantillon physique prélevé sur un hôte ou sur un organisme porteur (vecteur,

prédateur du vecteur). Une souche est un ensemble d'isolats ayant en commun plusieurs propriétés qui les caractérisent. Une espèce virale est donc constituée d'une ou plusieurs souches elle mêmes constituées d'un ou plusieurs isolats.

Individuals	Characters							
	a	b	c	d	e	f	g	h
1	+	+	+	−	−	−	−	−
2	+	+	+	−	−	−	−	−
3	+	+	−	+	−	−	−	−
4	+	+	−	+	−	−	−	−
5	−	−	−	−	+	+	+	−
6	−	−	−	−	+	+	−	+
7	−	−	−	−	+	−	+	+
8	−	−	−	−	−	+	+	+

Tableau SB.1: **Regroupement d'individus (1-8) selon des caractères (a-h) monothétiques ou polythétiques.** Les individus 1 à 4 peuvent être considérés comme un groupe monothétique car ils ont en commun les caractères a et b, de plus les individus 1 et 2 peuvent former un groupe monothétique car ils sont les seuls à partager les caractères a, b et c. De même les individus 3 et 4 peuvent former un groupe monothétique définit sur les caractères a, b et d qui leur sont communs. Les individus 5 à 8 forment un groupe polythétique car ils partagent 3 des 4 caractères e, f, g et h mais aucun de ces 4 caractères n'est partagé par tous les individus à la fois. D'après Gibbs *et al.*, 2006.

1.4.2. Le genre

Chaque genre est défini par un ensemble de caractéristiques communes des espèces qui la constituent. D'un point de vue général, les critères qui définissent le genre sont : l'organisation du génome, les gènes qu'il contient et le type de vecteur. Contrairement à l'espèce qui est définie selon un classement polythétique, un virus doit obligatoirement remplir tous les critères définissant le genre pour y appartenir (classement monothétique). Pour chaque genre, l'ICTV a défini une espèce-type qui est la plus connue ou la première décrite ou la mieux étudiée et qui partage l'essentiel des propriétés du genre. Il existe quelques espèces virales sans genre attribué mais qui peuvent tout de même appartenir à une famille. L'ICTV, en 2013, a répertorié 455 genres viraux.

1.4.3. La famille

Les genres sont eux-mêmes classés en familles selon un groupement monothétique sur la base de la structure du génome et la stratégie de réplication. L'ICTV répertorie 103 familles virales. Certains virus ne sont pas attribués à une famille, ainsi leur classification s'arrête au genre. On dénombre ainsi 17 genres qui ne sont pas assignés à une famille.

1.4.4. L'ordre

Les ordres regroupent des familles partageant une phylogénie commune. Beaucoup de familles virales ne sont pas assignées à un ordre (77/103), en effet, il existe seulement sept ordres actuellement répertoriés par l'ICTV (*Nidovirales*,

Mononegavirales, *Caudovirales, Herpesvirales, Ligamenvirales, Picornavirales, Tymovirales*).

1.5. Diversité et taxonomie des virus de plantes

1.5.1. Classification des virus de plantes

Les phytovirus sont répartis de manière ubiquiste dans le règne végétal, ils ont été décrits à partir d'algues, de gymnospermes, de ptéridophytes, de bryophytes et d'angiospermes (Hull, 2002). Concernant leur nomenclature, le nom de l'espèce pour les phytovirus est généralement composé du nom de l'hôte à partir duquel le premier membre a été décrit, du type de symptômes qu'il induit, suivi parfois de l'origine géographique de sa détection, puis du terme 'virus' (exemples : *Barley yellow dwarf virus* (BYDV), *Tomato leaf curl New Delhi virus* (ToLCNDV)). Concernant le nom du genre il est composé des premières lettres des mots qui désignent l'espèce représentative auquel on ajoute le suffixe 'virus' (par exemple le genre Mastrevirus provient de la contraction de ***Ma****ize* ***stre****ak virus* (MSV)). Les familles et les ordres sont nommés avec le suffixe 'viridae' et 'virales', respectivement.

Le dernier rapport de l'ICTV datant de juillet 2013 et mis à jour le 30 juin 2014 répertoriait 1210 espèces de virus de plantes réparties dans 113 genres dont 23 non assignés à une famille, et 20 familles (Figure SB.5et Annexe 1). Il est important de souligner le fait que certaines familles virales contiennent des virus pouvant infecter des hôtes de différents règnes. Ainsi, la famille *Rhabdoviridae* contient à la fois des virus de plantes et des virus d'animaux.

1.5.2. Une diversité phytovirale sous-estimée et biaisée

Il est difficile d'estimer le nombre exact de phytovirus actuellement découverts et décrits. En effet, l'ICTV ne répertorie que les virus qui leur ont été soumis via un formulaire de proposition d'espèce. Il existe donc un nombre indéfini de phytovirus dont les séquences sont présentes dans des bases de données telles que la GenBank de NCBI (National Center for Biotechnology Information) qui ne sont pas répertoriées par l'ICTV. La production massive de données telles que des « reads » (lecture de séquences), « contigs » (assemblage de séquences qui partagent des régions communes), « scaffolds » (assemblage de contigs), génomes assemblés *de novo*, etc., est destinée à augmenter avec l'avènement des nouvelles technologies de séquençage (NGS). L'accumulation de ces nouvelles données incomplètes ne permet plus d'avoir une vision synthétique et actualisée de la diversité des virus en général, et des phytovirus en particulier.

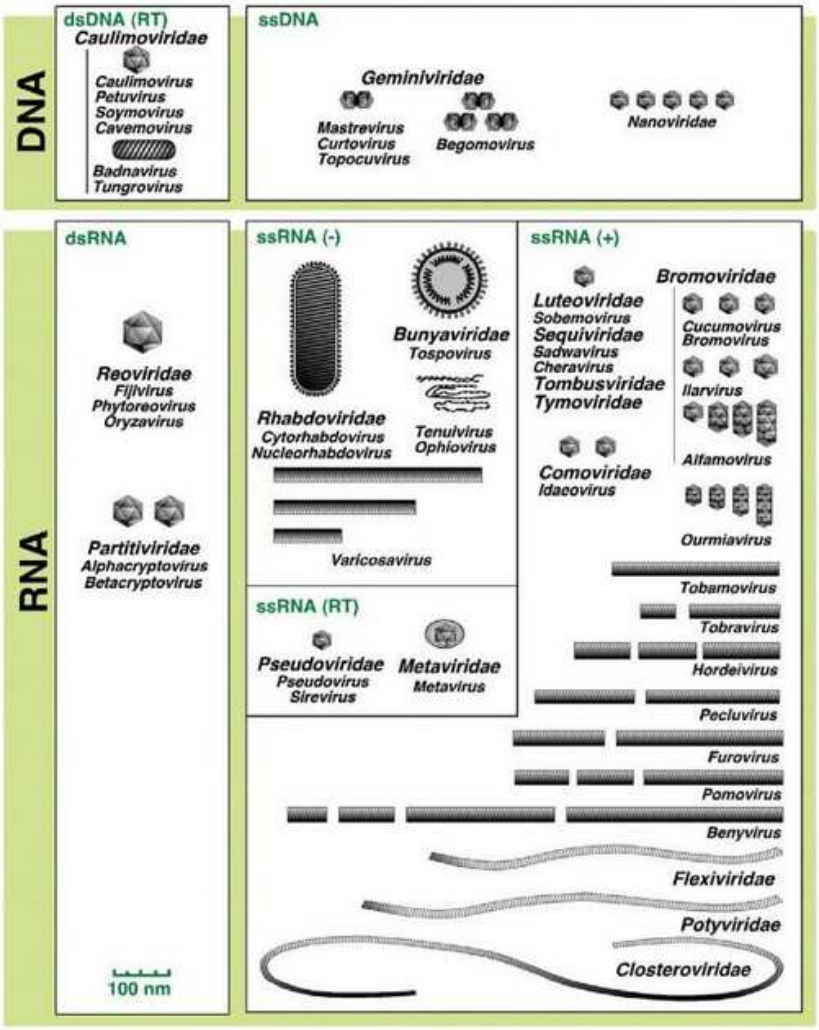


Figure SB.5: Représentation de la diversité de formes des virions de plantes (ce tableau n'est pas exhaustif). D'après Cann, 2012.

En 2006 l'ICTV répertoriait 766 espèces phytovirales, puis 900 en 2009 et 1210 en 2013. Cette augmentation semble linéaire, contrairement au volume des séquences virales contenu dans les bases de données qui semble augmenter de manière exponentielle (Figure SB.6). On peut donc penser qu'on assiste de nos jours à un « décrochage » entre deux types de diversités, la première de nature « encyclopédique » qui récence les espèces virales par des outils classiques de virologie et la deuxième « virtuelle » basés sur des données de séquences, souvent incomplètes, et contenant potentiellement de nouvelles espèces.

On peut donc penser que la vision que nous avons de la diversité phytovirale est sous-estimée. Par ailleurs, elle est probablement biaisée comme l'ont montré plusieurs études de revue (Cooper and Jones, 2006; Roossinck, 2011a; Wren *et al.*, 2006). Globalement, ce biais peut être expliqué par trois facteurs majeurs : premièrement, la virologie végétale est un domaine où les financements sont moins abondants qu'en virologie humaine/animale, ce qui réduit donc le nombre d'études se focalisant sur les virus « verts »; deuxièmement la majorité des études sur les virus de plantes se sont

focalisées sur les plantes d'intérêt agronomique (Wren *et al.*, 2006) (Figure SB.7) laissant de côté les virus des plantes sauvages ou non-cultivées (adventices, haies, friches etc.), troisièmement la découverte de phytovirus est souvent conditionnée à l'observation de symptômes sur une plante hôte, alors qu'il a été récemment suggéré que seule une faible fraction des virus de plantes causeraient des dommages à leur hôte (Roossinck, 2005).

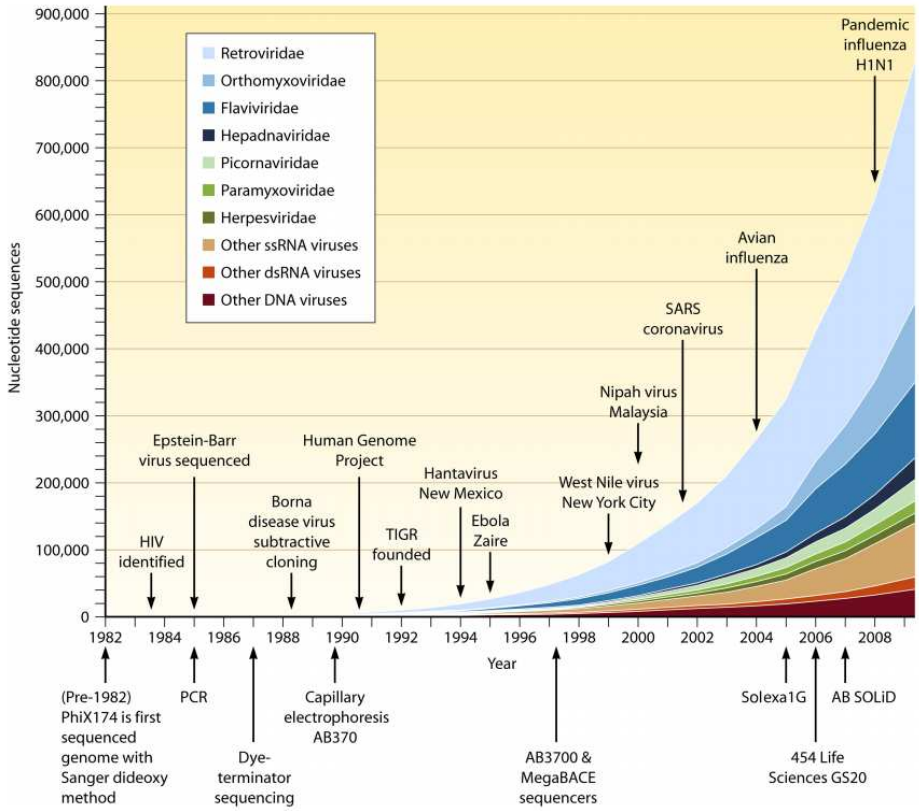


Figure SB.6: Croissance des bases de données de séquences virales depuis 1982. D'après Lipkin, 2010.

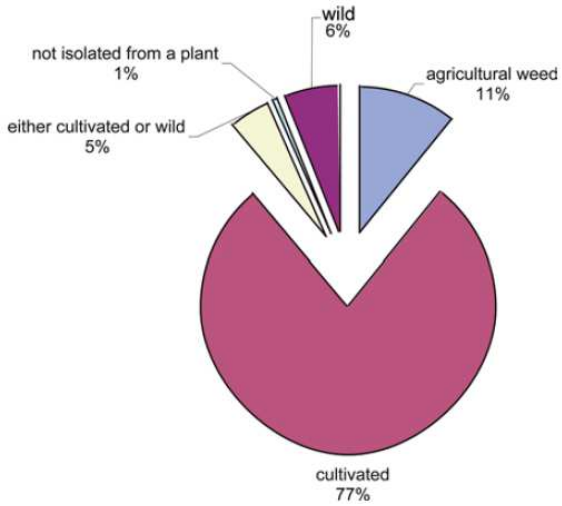


Figure SB.7: Fréquence des plantes sources à partir desquelles les espèces de phytovirus ont été initialement isolées. D'après Wren *et al.*, 2006.

1.5.3. Les écosystèmes sauvages : une terre encore vierge pour la virologie végétale ?

L'ensemble de ces résultats et hypothèses récentes nous amène donc à supposer que la majorité des études menées jusqu'à nos jours auraient sous-estimé le rôle potentiel du milieu sauvage en terme de réservoir de biodiversité phytovirale. Quelques études d'envergure visant à décrire la diversité des virus associés aux plantes sauvages ont été effectuées ces dernières années (Muthukumar *et al.*, 2009; Roossinck *et al.*, 2010). Ces travaux semblent indiquer que des milliers de virus de plantes restent à découvrir ; par exemple une étude de Marylin J. Roossinck menée sur 7000 espèces de plantes sauvages au Costa Rica en 2010 a démontré que la présence de virus dans les plantes était un phénomène commun. En effet, des virus, ou plus précisément des traces – reads/contigs, ont été détecté au niveau de 70% des plantes analysées (Roossinck, 2012b; Roossinck *et al.*, 2010; Wren *et al.*, 2006) (cf. section 2.3).

1.6. Composantes et mesures de la diversité dans les communautés

1.6.1. Composantes de la diversité

On distingue plusieurs composantes permettant de décrire la diversité en espèces d'une communauté : la richesse, l'équitabilité la régularité et la divergence.

- La richesse est le nombre de classes (espèce, genre, famille, ordre) présentes dans le système étudié. Cette définition assume deux hypothèses : les classes sont bien définies et elles sont équidistantes (c'est-à-dire que la richesse augmente d'une unité lorsqu'on rajoute une espèce, que cette espèce soit proche ou distante des autres). L'indice de richesse le plus utilisé et le plus simple est le nombre d'espèces *S*.

- L'équitabilité représente la régularité de la distribution des espèces. Elle tient compte du fait qu'une espèce représentée abondamment ou par un seul individu n'apporte pas la même contribution à l'écosystème. Par exemple, en considérant que le nombre total d'individus dans une communauté est limité, si une espèce est dominante, elle entrainera, mathématiquement, la rareté d'autres espèces. Les indices d'équitabilité les plus couramment utilisés sont les indices de Simpson (Simpson, 1949) et de Shannon-Wiener (Shannon, 1948) qui évaluent à la fois la richesse et l'équitabilité.

- La régularité et la divergence tiennent compte du fait que deux espèces du même genre sont plus proches que deux espèces de familles différentes. Cette notion est implémentée dans les mesures de diversité phylogénétique et de diversité fonctionnelle. Les mesures de divergences sont construites à partir de la dissimilarité entre classes avec ou sans pondération par la fréquence. Les mesures de régularité décrivent quant à elles la façon dont les espèces occupent l'espace des niches (régularité fonctionnelle). Ce concept complète celui d'équitabilité dans les mesures classiques de diversité : la diversité augmente avec la richesse, l'équitabilité et la régularité.

1.6.2. Les différents niveaux de diversité

La diversité peut être décrite à plusieurs niveaux, on parle de diversité α, β et γ (Whittaker, 1960) (Figure SB.8).

La diversité α représente la diversité locale qui va être mesurée à l'intérieur d'un habitat uniforme et de taille fixe.

La diversité β consiste à comparer la diversité des espèces entre habitats ou écosystèmes ou le long de gradients environnementaux. Cet indice donne une indication de la variation en composition d'espèces dans l'ensemble des habitats ou parmi des communautés.

La diversité γ mesure la diversité sur la totalité du système étudié, c'est-à-dire sur la totalité des habitats pris en compte. Sa mesure s'effectue donc de la même manière que celle de la diversité α (c'est-à-dire via les mêmes indices de diversité). La diversité α sera donc plutôt définie au niveau de la communauté, alors que la diversité γ sera définie au niveau de la méta-communauté.

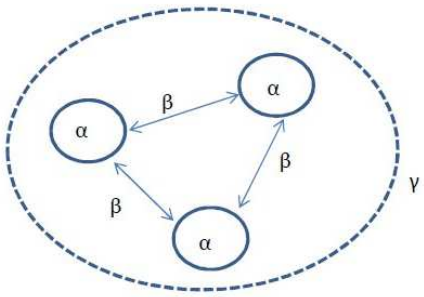


Figure SB.8: Représentation des trois niveaux de description de la diversité. Source : http://ideas4sustainability.wordpress.com/2012/03/22/the-richness-of-diverse-meanings/.

1.6.3. Indices de diversité

Il existe un large panel d'indices permettant de décrire la diversité, ici nous ne traiterons que de ceux utilisés le plus fréquemment.

1.6.3.1. Mesure de la diversité α et γ

- ***Indice de Simpson***

Cet indice mesure la probabilité que deux individus sélectionnés au hasard n'appartiennent pas à la même espèce. Il tient compte de l'abondance et de la richesse en espèces de la communauté. L'indice de Simpson est sensible aux variations d'importance des espèces les plus abondantes.

$$Indice\ de\ Simpson, D = 1 - \sum_{i}^{S} p_i^2$$

S est le nombre total d'espèces dans la communauté, c'est-à-dire la richesse spécifique, P_i est la proportion d'individus de chaque espèce i au sein de l'échantillon. L'indice de Simpson dépend donc de la richesse spécifique et de l'équitabilité. L'indice varie entre 0 quand l'échantillon est constitué d'une seule espèce et tend vers 1 avec l'augmentation du nombre d'espèces.

- ***Indice de Shannon-Wiener***

Tout comme l'indice de Simpson, l'indice de Shannon-Wiener exprime la diversité d'une communauté en tenant compte de l'abondance et de la richesse en espèces. A la différence de l'indice de Simpson, celui de Shannon donne un « poids » à l'abondance, ainsi, le « poids » des espèces abondantes va être légèrement réduit comparé aux espèces rares.

$$Indice\ de\ Shannon - Wiener, H' = -\sum_{i=1}^{S} P_i \ln P_i$$

Une communauté présentant une espèce dominante aura un indice H plus petit qu'une communauté dans laquelle les espèces ont une abondance proche. L'indice varie entre 0 quand la l'échantillon est constitué d'une seule espèce et tend vers lnS lorsque toutes les espèces ont la même abondance.

1.6.3.2.Mesure de la diversité β

- ***Indices de similarité de Sørensen et de Jaccard***

Ces deux indices sont basés sur le même principe dans le sens où ils tiennent compte du nombre d'espèces communes entre les 2 communautés comparées.

$$Indice\ de\ Sørensen, L = \frac{2C}{2C + A + B}$$

$$Indicede\ Jaccard, J = \frac{C}{C + A + B}$$

C est le nombre d'espèces communes aux deux communautés, A le nombre d'espèces qui est unique à la communauté 1 et B le nombre d'espèces qui est unique à la communauté 2. Ces indices varient de 0, quand aucune espèce n'est commune entre les deux communautés, à 1 lorsque ces communautés sont composées des mêmes espèces.

- ***Indice de Morisita-Horn***

Les indices de Jaccard et de Sørensen sont les plus utilisés, cependant ils sont très sensibles à la taille de l'échantillon, surtout quand ces derniers contiennent de nombreuses espèces rares, de plus, il ne tiennent pas compte de l'abondance des individus au sein de l'espèce (ils ne se basent que sur la présence/absence)(Chao *et al.*, 2004; Chao *et al.*, 2006). Ces mesures sous-estimeraient ainsi la similarité entre deux communautés. Afin de palier à ces problèmes, l'indice de Morisita-Horn est considéré comme un des indices les plus robustes et fiables (Chao *et al.*, 2006; Wolda, 1981). En

effet, cet indice correspond au rapport de la probabilité que deux individus tirés au hasard dans 2 échantillons différents appartiennent à la même espèce sur la probabilité que 2 individus tirés au hasard dans le même échantillon appartiennent à la même espèce. Par conséquent, son calcul ne va pas être influencé par la taille de l'échantillon, et il va tenir compte de l'abondance des individus à l'intérieur de chaque espèce.

2. Ecologie virale : de l'organisme au paysage

Les maladies infectieuses sont représentées schématiquement dans un triangle comprenant les interactions hôte-agent pathogène-environnement (Figure SB.9a). En effet, pour qu'une maladie se développe, l'agent pathogène qui en est responsable doit être virulent et il a besoin d'un hôte sensible et d'un environnement favorable. Si une des trois parties de ce triangle est absente, il n'y aura pas de maladie (Figure SB.9bc). Ce triangle est conceptuellement à la base de la phytopathologie.

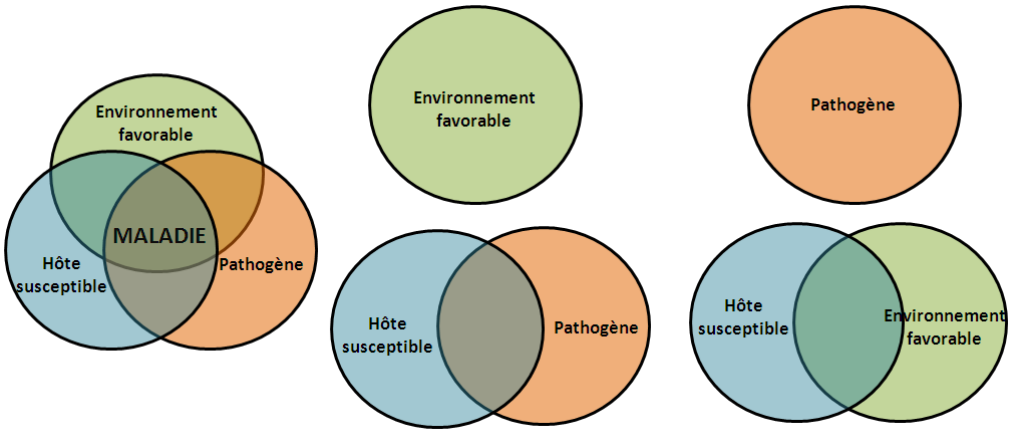


Figure SB.9: **Le triangle épidémiologique**. a) Tous les acteurs du triangle sont réunis pour provoquer la maladie b) Au moins un des trois facteurs est absent et la maladie ne se développera pas.

Bien que l'écologie virale soit née dans la première moitié du 20e siècle (Malmstrom *et al.*, 2011), la majorité des connaissances que nous avons sur les interactions « hôtes-organismes phytopathogènes-environnement » proviennent de l'étude des champignons pathogènes des plantes (Alexander, 2010; Alexander *et al.*, 2013; Burdon, 1987; Gilbert, 2002). À ce jour, peu d'études ont été effectués en écologie virale.

2.1. Les interactions plantes/virus à l'échelle des individus

2.1.1. La coévolution plante/virus : résistance et pathogénie

De manière générale, les virus altèrent l'expression des gènes et la physiologie de leur plante hôte, ayant des effets directs sur la survie de l'hôte, sa croissance, sa fécondité et par conséquent sur sa valeur sélective ou fitness (c'est-à-dire sa capacité à survivre et à se reproduire) (Gilbert, 2002). De plus, les changements physiologiques induits par l'infection virale dans la plante hôte peuvent être influencés par la disponibilité des ressources pour la plante. Les effets des virus sur les plantes existent

au sein d'un continuum allant du positif au négatif, de plus, les conséquences de l'infection vont être modulées par des facteurs biotiques et abiotiques (Alexander *et al.*, 2013). Selon que l'on se situe dans le milieu cultivé ou sauvage le questionnement ne sera pas le même vis-à-vis de l'effet des virus sur leur hôte. En agronomie on se demandera si le virus affecte le rendement alors qu'au niveau des plantes sauvages on se posera des questions sur la fitness de la plante hôte (Alexander *et al.*, 2013), bien que le rendement puisse intervenir dans la fitness.

La définition de la coévolution sur laquelle nous nous basons ici est la suivante : changement évolutif d'un trait chez une espèce en réponse à un trait chez une deuxième espèce, suivie d'une adaptation évolutive de cette deuxième espèce (Janzen, 1980). Par conséquent, la coévolution décrit les transformations qui se produisent au cours de l'évolution entre deux ou plusieurs espèces suite à leurs influences réciproques. Ainsi, lorsqu'un agent pathogène réduit la fitness de l'hôte, il exerce une pression de sélection sur son hôte qui peut conduire au développement de différentes stratégies de défense qui limitent l'infection (Agnew *et al.*, 2000). Chez les plantes, il existe deux stratégies de défense : la résistance qui est la capacité de l'hôte à limiter la multiplication de l'agent pathogène et la tolérance qui est la capacité de l'hôte à réduire les dommages induits par l'infection (Clarke, 1986).

2.1.1.1. Les mécanismes de défense chez les plantes impliqués dans la résistance

Les plantes n'ont pas de système immunitaire comparable à celui des vertébrés. Le système de défense des plantes est constitué de plusieurs niveaux d'action qui leur permet d'éviter, de supprimer, ou de se défendre activement contre les agents pathogènes (de Ronde et *al.* 2014). Un virus ne sera capable de provoquer une infection que dans les plantes hôtes qui n'auront pas développé de défenses spécifiques en réponse aux facteurs de virulence (Pallas and Garcia, 2011).

- ***La résistance non-hôte***

Les mécanismes de résistance non-hôte (NHR) (Uma *et al.*, 2011) sont des mécanismes de résistance génériques et non spécifiques. Parmi les NHR on va distinguer les mécanismes de défense basiques qui empêchent l'invasion par l'agent pathogène (production de métabolites secondaires, épaississement de la paroi cellulaire, etc.) et la réponse hypersensible (HR) qui consiste à confiner l'agent pathogène dans son aire d'inoculation (via la mort programmée des cellules végétales) de manière à empêcher sa propagation (Soosaar *et al.*, 2005).

- ***Le RNA silencing***

Le RNA silencing ou RNA interference (RNAi) est une sorte de système immunitaire inné de la plante : la cellule hôte possède une machinerie permettant de cliver les structures d'ARN double brin (formes réplicatives de certains virus) via des complexes appelés « dicers », ce qui va amener à la formation de petits ARNs

interférents (siRNA) et microRNA (miRNA) de 21 à 24 nucléotides qui vont eux même être chargés dans des complexes d'effecteurs RISC contenant une protéine Argonaute (AGO). Ces mécanismes se déroulent à deux niveaux. Au niveau post-transcriptionnel, les RISC dégradent les ARNs viraux ou inhibent leur traduction. Les complexes RISC ont également une action au niveau transcriptionnel via des modifications de l'ADN ou des histones menant à la formation d'hétérochromatine permettant ainsi l'inhibition de la synthèse d'ARN (Baulcombe, 2004; Voinnet, 2001) (Figure SB.10). Les virus sont la cible de ce système quand le génome ou les ARNs messagers viraux, notamment par des structures secondaires, présentent une forme en double brin.

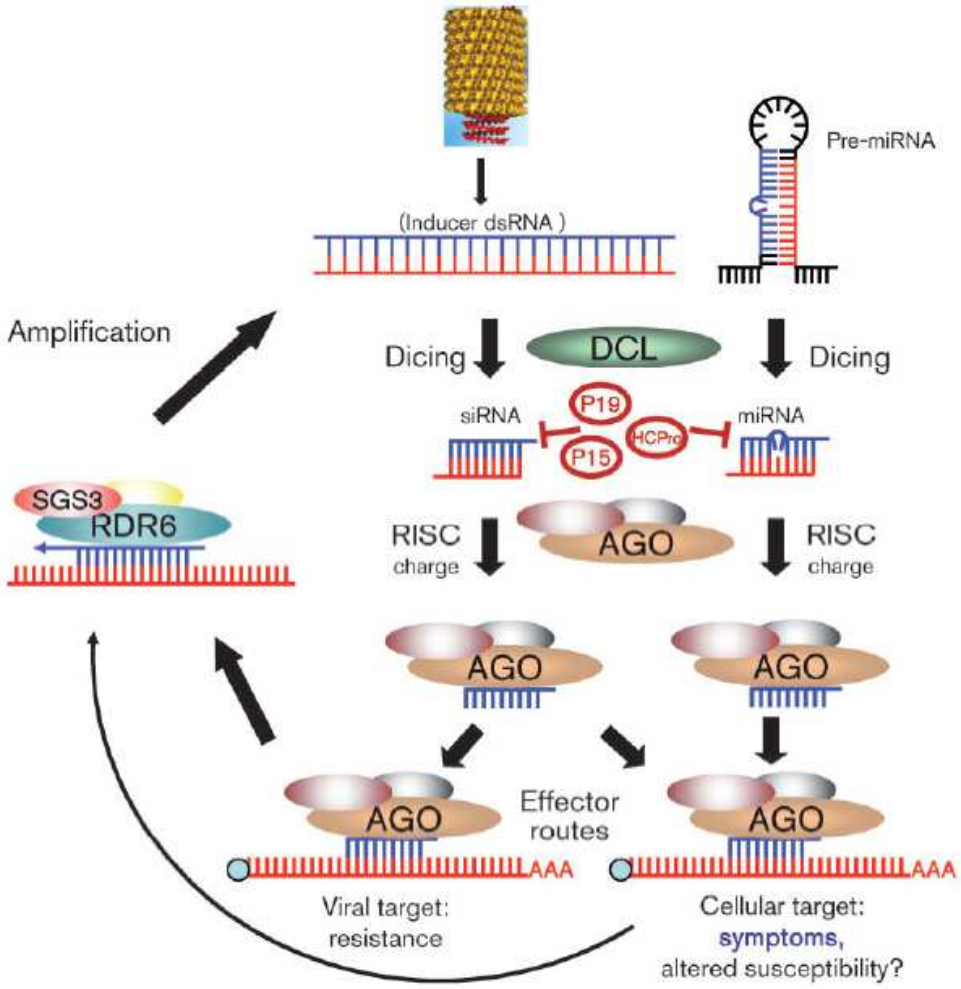


Figure SB.10 : Schéma des étapes clé du RNA silencing. Les molécules de dsRNA et les premiRNAs cellulaires sont coupés par des RNases Dicer-like (DCL), générant ainsi des siRNAs et des miRNAs respectivement. Ces siRNAs et miRNAs sont chargés dans des complexes d'effecteurs RISC contenant des protéines Argonaute (AGO) et dirigés vers des cibles homologues. Les produits générés par les RISC vont être incorporés dans un cycle d'amplification médié par la polymérase RDR6 et la protéine SGS3 afin de générer des siRNAs. D'après Pallas and Garcia, 2011.

- ***Les interactions gène-pour-gène et la résistance quantitative***

La coévolution dans les systèmes plantes-agents pathogènes a souvent été analysée dans le cadre des interactions gène-pour-gène, sa définition étant que pour chaque gène d'avirulence (*avr*) de l'agent pathogène, il y a un gène correspondant (*res*)

codant pour la résistance de la plante hôte (Figure SB.11). Le contournement des systèmes de résistance gène-pour-gène peut conduire à l'émergence de maladies virales. À titre d'exemple, on peut citer le cas du *Bean common mosaic virus* (BCMV) infectant le haricot *Phaseolus vulgaris*. Sept souches de BCMV présentent une interaction gène-pour-gène avec différents génotypes de haricots, ces souches vont alors pouvoir émerger lorsqu'elles perdent des gènes d'avirulence codant pour des protéines d'avirulence (AVR) reconnues par les protéines de résistance de la plante hôte (RES) (Anderson *et al.*, 2004). Bien que ces résistances soient qualitatives (c'est-à-dire que la plante est soit sensible soit résistante), dans plusieurs cas elle peut être quantitative et/ou polygénique et héréditaire (Fraile and Garcia-Arenal, 2010; Maule *et al.*, 2007). Une résistance quantitative va induire des différences dans les phénotypes résistants (allant du plus sensible au plus résistant) (St Clair, 2010). Par exemple, chez le piment, 11 régions chromosomiques ont été associées à une résistance quantitative au *Potato virus Y* (PVY), cette résistance résulte d'une association d'un QTL (Quantitative Trait Loci) à effet majeur avec plusieurs QTLs à effets mineurs (Caranta *et al.*, 1997).

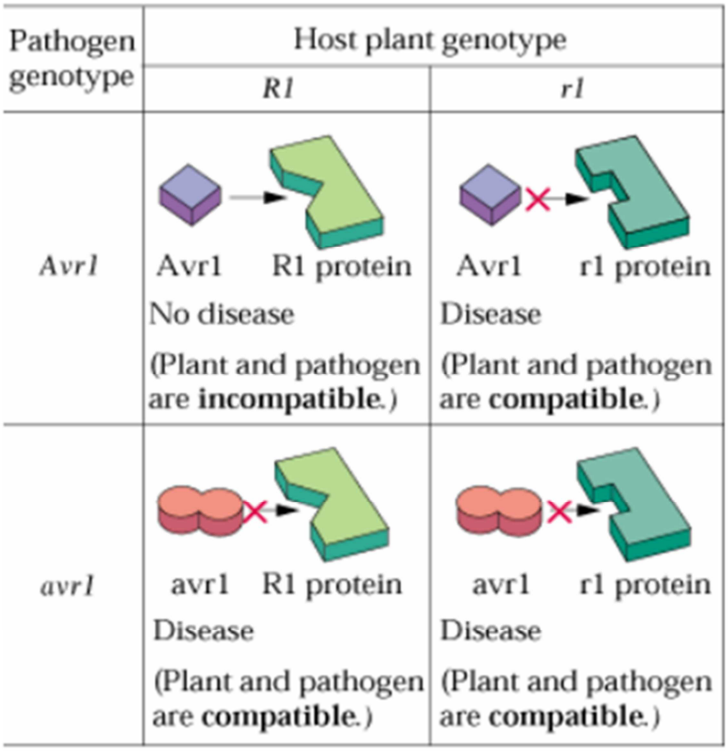


Figure SB.11 : Illustration de la théorie « gène pour gène » dans les interactions plantes-agents pathogènes. Lorsque la protéine (Avr1) issue du gène d'avirulence de l'agent pathogène est reconnue par la protéine (R1) issue du gène de résistance de la plante la maladie ne se développe pas. La maladie se développera dans le cas où la protéine d'avirulence n'est pas reconnue par la protéine de résistance. Source : https://www.bordeaux.inra.fr/umr619/telechargement/GDP802-VL-2007.pdf.

2.1.1.2. La tolérance : une stratégie de défense alternative

Les mécanismes impliqués dans la tolérance ne sont pas bien compris, en effet, l'étude de la tolérance a reçu beaucoup moins d'attention de la part des scientifiques comparé à la résistance (Fraile and Garcia-Arenal, 2010). Il semblerait que ces

mécanismes soient reliés à la capacité de la plante à modifier ses propres traits d'histoire de vie face à l'infection (Pagan *et al.*, 2008). Le contrôle de la tolérance serait polygénique et parfois monogénique (Clarke, 1986). Ainsi, la tolérance dépendrait de l'interaction entre les génotypes des plantes et des virus (Pagan *et al.*, 2008). Bien que la tolérance ne soit pas directement antagoniste à la fitness des pathogènes, l'évolution vers la tolérance peut amener à des changements sélectifs des populations d'agents pathogènes provoquant ainsi leur évolution vers un contournement de la tolérance (Little *et al.*, 2010).

2.1.1.3. Résistance et pathogénie dans les systèmes agricoles vs. sauvage

Dans le cadre de la coévolution plante-agents pathogènes vis-à-vis des interactions gène-pour-gène, deux théories majeures sont invoquées (Brown and Tellier, 2011) :

(1) La théorie de la Reine Rouge ou course aux armements : la virulence d'un agent pathogène va amener à une résistance accrue de la plante hôte, cette augmentation de la résistance va par la suite mener à une augmentation de la virulence de l'agent pathogène et ainsi de suite. Dans ce cas, les allèles des gènes de résistance et d'avirulence se succèdent rapidement et amènent par conséquent à une situation dans laquelle les allèles des gènes *res* et *avr* vont être temporairement fixés (Bergelson *et al.*, 2001).

(2) La théorie de la guerre des tranchées : les allèles de résistance et d'avirulence ont un polymorphisme balancé et changent très peu au cours du temps dans le sens où ils peuvent « avancer » mais aussi « battre en retraite », la situation est alors considérée comme stable dans le temps (Stahl *et al.*, 1999; Woolhouse *et al.*, 2002). Dans ce cas, il va y avoir une persistance à long terme de la variation génétique des allèles des gènes *res* et *avr*.

La majorité des preuves de la coévolution entre les plantes et leurs agents pathogènes provient de l'étude de systèmes agricoles dans lesquels les pathogènes sont fortement virulents et les populations végétales sont génétiquement homogènes (Fraile and Garcia-Arenal, 2010). Ainsi, l'hypothèse selon laquelle les virus diminuent la fitness de leurs hôtes est majoritairement basée sur la sévérité des symptômes et la baisse de productivité qu'ils induisent sur les plantes cultivées (Fraile and Garcia-Arenal, 2010). Dans les systèmes agricoles, l'Homme sélectionne des cultivars génétiquement homogènes et avec une résistance accrue aux agents pathogènes. Ces cultivars exercent une forte pression de sélection sur l'agent pathogène en faveur d'une augmentation de la virulence et finalement d'un contournement de la résistance. Un nouvel effort de sélection est nécessaire pour produire des cultivars résistants à ces nouvelles souches virulentes selon la théorie de la course aux armements (Brown and Tellier, 2011). Par ailleurs, il existe des preuves démontrant que les gènes *res* et *avr* ont évolué rapidement au cours du temps dans les systèmes agricoles depuis la domestication des plantes (reporté dans la revue de Brown and Tellier, 2011).

En revanche, il existe peu de preuves démontrant que les virus en milieu sauvage ont un effet négatif sur la fitness de leurs plantes hôtes (Fraile and Garcia-Arenal, 2010). Le système naturel le mieux documenté à ce jour est celui des graminées sauvages infectées par le *Barley yellow dwarf virus* et le *Cereal yellow dwarf virus* (B/CYDVs) (Malmstrom *et al.*, 2006; Power and Mitchell, 2004). Dans ce cas précis, l'infection virale peut avoir un effet négatif direct et indirect sur la fitness des hôtes (cf. section 2.2.3). Il a par ailleurs été documenté chez l'avoine sauvage en Australie une structuration des génotypes résistants en corrélation positive avec les génotypes virulents de champignons phytopathogènes : les génotypes végétaux les plus résistants sont associés aux champignons qui sont les plus virulents (Burdon and Thrall, 2008). Toutefois, l'hypothèse majoritairement invoquée en milieu sauvage est celle de la guerre des tranchées. En effet, dans la nature, les gènes *res* et *avr* ont une durée de vie longue qui est cohérente avec l'idée que ces gènes aient pu avoir une fonction propre et indispensable durant des milliers voire des millions d'années (Bergelson *et al.*, 2001; Stahl *et al.,* 1999). De plus, les plantes sauvages sont soumises à une forte variabilité de leur habitat et de leur structuration géographique ce qui amènerait à un grand polymorphisme de leurs allèles de résistance.

Cette dernière décennie, l'idée selon laquelle la majorité des infections virales en milieu sauvage ne provoqueraient pas de symptômes a émergé (Cooper and Jones, 2006; Prendeville *et al.*, 2012). Ceci a amené à un questionnement quant aux types d'interactions plantes-agents pathogènes qui existent en milieu sauvage : sont-ils forcément de type antagoniste ? Dans ce sens, il a été proposé que dans certaines conditions, des virus pourraient avoir des interactions de type mutualiste avec leurs plantes hôtes (Roossinck, 2005; Wren *et al.*, 2006).

2.1.1.4. Une entorse à la règle générale : un virus peut-il être mutualiste ?

Depuis les années 2000, Marylin J. Roossinck a fait émerger le concept de virus mutualiste. En effet, les études de métagénomique qu'elle a développées ont permis de montrer que la majorité des virus en milieu sauvage ne semblent pas induire de symptômes apparents (Roossinck, 2012b; Roossinck *et al.*, 2010). L'hypothèse selon laquelle certains virus pourraient être bénéfiques pour leur hôte a alors été proposée (Roossinck, 2005; Roossinck, 2010; Roossinck, 2011b; Roossinck, 2011c; Roossinck, 2013). Cette hypothèse a été vérifiée en laboratoire pour quelques systèmes plante-virus. Ainsi, le *Cucumber mosaic virus* confère la résistance au froid aux betteraves (Xu *et al.*, 2008). De même, l'association du Curvularia thermal tolerance virus avec le champignon endophyte *Curvularia protuberata* permet à la plante *Dichanthelium lanuginosum* de croître à des températures élevées (Marquez *et al.*, 2007). Les mécanismes de ces interactions bénéfiques restent cependant encore méconnus (Roossinck, 2013). Il semble que les effets positifs des virus sur leur plante hôte ne sont probablement pas dus à une interaction directe du virus avec la plante hôte. En effet, on peut citer l'étude d'Adrian Gibbs en 1980 dans laquelle il a démontré que des légumes natifs australiens inoculées avec un *Tymovirus* avaient une taille réduite ce qui leur

permettait d'être moins pâturés que les légumes sains, il y a donc un effet positif indirect de ce virus sur la survie de son hôte (Gibbs, 1980).

Cependant, il faut être conscient du fait que l'effet des virus sur un individu ne permet pas de prédire leur effet sur la communauté (Alexander *et al.,* 2013).

2.1.2. Transmission et gamme d'hôte

2.1.2.1. La transmission

La transmission est un évènement obligatoire pour la survie des virus et il existe un nombre varié de mécanismes de dissémination des phytovirus (Figure SB.12).

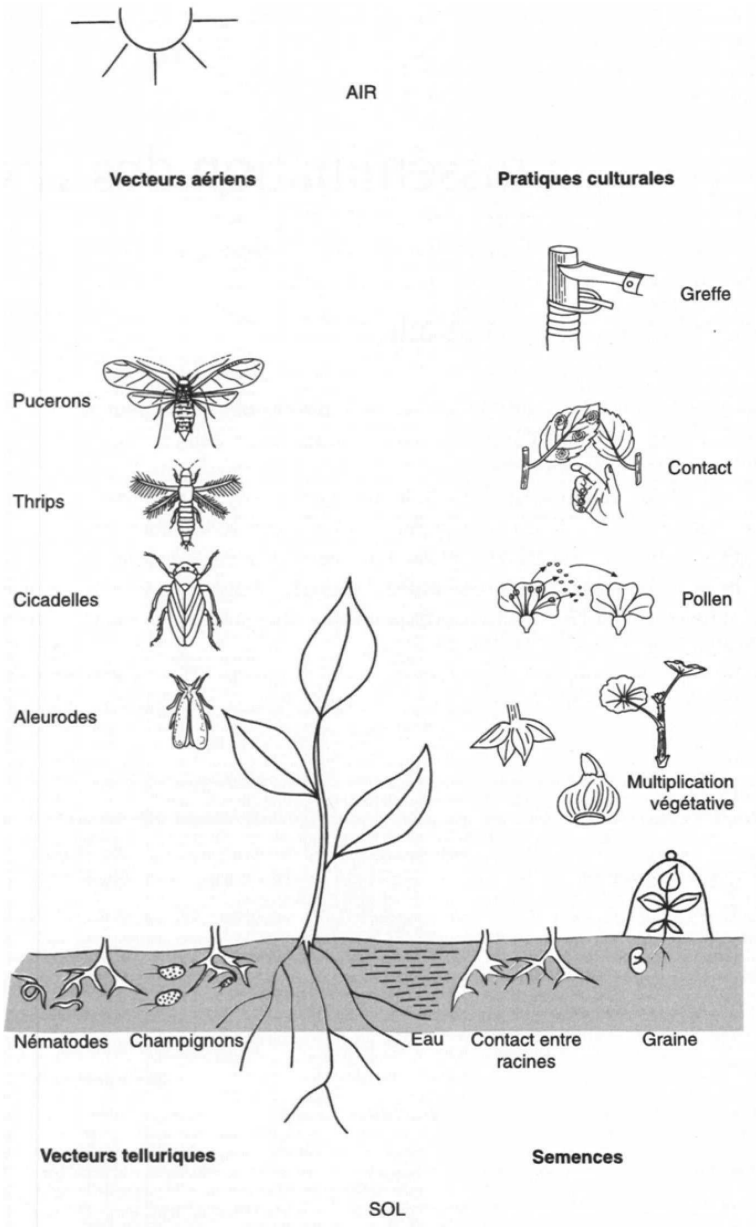


Figure SB.12 : Diversité des modes de dissémination des phytovirus : transmission par les organes de multiplication végétative ou par la graine (transmission verticale), transmission par contact ou vecteurs aériens ou telluriques (transmission horizontale). D'après Astier, 2007.

- ***Transmission verticale***

Certains phytovirus se transmettent à la descendance de la plante infectée par la graine ou par multiplication végétative, on parle alors de transmission verticale. Parmi les virus transmis verticalement on trouve les virus définis récemment comme étant persistants (Roossinck, 2010; Roossinck, 2012a). Ces virus, que l'on retrouve dans les familles *Endornaviridae et Partitiviridae* subsistent dans leur plante hôte à long terme et sont transmis à des taux avoisinant les 100% (Roossinck, 2010; Roossinck, 2013). Par ailleurs, une nouvelle famille de virus proche des *Partitiviridae*, les *Amalgaviridae*s, sont aussi fortement soupçonnés d'être des virus persistants (Liu *et al.*, 2012a; Martin *et al.*, 2011c). L'étude d'écogénomique de Marylin J. Roossinck a par ailleurs démontré que les virus persistants représenteraient la majorité des virus de plantes sauvages avec des incidences allant jusqu'à 70% pour les *Partitiviridae* pour certaines familles botaniques (Roossinck, 2012a; Roossinck, 2012b). On peut tout de même trouver des virus persistants appartenant aux familles *Endornaviridae* et *Partitiviridae* sur quelques plantes cultivées telles que le riz (*Oryza sativa endornavirus*), le haricot *(Phaseolus vulgaris endornavirus*) ou le radis (*Raphanus sativus cryptic virus*) (reporté dans la revue de Roossinck, 2012a).

D'autres virus peuvent être transmis par la graine sans pour autant être persistants, ils sont largement distribués dans diverses familles de phytovirus (*Potyviridae, Virgaviridae, Tymoviridae, Secoviridae, Luteoviridae, Bromoviridae, Tombusviridae,* etc.) et vont engendrer des dommages sur leurs hôtes (Sastry, 2013). Parmi eux, on peut citer le cas largement étudié du *Pea seed-borne mosaic virus* (Potyvirus) qui cause de sérieux dommages aux légumineuses engendrant ainsi des pertes économiques considérables (Khetarpal and Maury, 1987).

- ***Transmission horizontale***

La majorité des virus de plantes (environ 90%) (Mink, 1993) sont transmis horizontalement par des vecteurs qui leur permettent de se disséminer sur de nouveaux individus hôtes. On trouve des vecteurs de virus de plantes chez les arthropodes (insectes, acariens), nématodes, et champignons (Astier, 2007; Bragard *et al.*, 2013). La transmission horizontale peut également avoir lieu via la greffe (Astier, 2007) ou encore par contact direct ou l'utilisation d'outils contaminés mais ces deux dernières possibilités n'ont été observées que dans des cas restreints (Tobamovirus, Potexvirus, et Hordeivirus) (Astier, 2007). Les vecteurs permettent ainsi aux virus de se disséminer sur des distances plus ou moins longues, par exemple les pucerons peuvent parcourir des centaines de kilomètres quand ils sont portés par le vent (Dixon, 1985).

- ***Vection***

Il existe une interaction étroite entre les vecteurs et les virus qu'ils transmettent. On distingue ainsi deux modes de transmission :

- La transmission dite circulante où le virus suite à son ingestion par l'insecte va atteindre l'intestin, traverser la paroi intestinale puis se diffuser dans l'hémolymphe pour atteindre les glandes salivaires. Lors de ce type de transmission le virus peut se répliquer ou non dans le vecteur (Bragard *et al.*, 2013).

- La transmission dite non circulante où le virus est retenu au niveau des stylets et/ou du tube digestif antérieur du vecteur. Ce type de transmission est divisé en deux sous-catégories : la transmission non-persistante pour laquelle le virus ne survit que sur une courte durée (quelques secondes à quelques minutes) dans l'insecte contrairement à la transmission semi-persistante (l'insecte reste virulifère durant quelques heures à quelques jours) (Bragard *et al.*, 2013).

- ***Facteurs écologiques influençant la transmission***

Des caractéristiques intrinsèques aux insectes vecteurs vont influencer l'écologie des phytovirus (Power, 2008). Parmi ces caractéristiques on peut citer la préférence d'hôte ; la gamme d'un virus obligatoirement transmis par vecteur est forcément limitée par la gamme d'hôte du vecteur (Power, 2008). De plus, la probabilité qu'un évènement de transmission se produise par unité de temps dépend inévitablement de la densité de la population de vecteur et de son comportement (Madden *et al.*, 2000). Cependant, le virus peut dans certains cas manipuler son vecteur afin d'augmenter sa probabilité d'être transmis (Desbiez *et al.*, 2011; Elena *et al.*, 2014). Par exemple, plusieurs études ont montré que la fécondité des vecteurs était accrue sur les plantes infectées par des virus (reporté dans la revue de Fereres and Moreno, 2009). De plus, une étude sur le *Caulimoflower mosaic virus* (CaMV) a démontré qu'un virus peut « percevoir » la présence de son vecteur et ainsi augmenter la probabilité d'acquisition. En effet, le corps à transmission du CaMV réagit instantanément à la présence du puceron vecteur par une redistribution rapide et réversible de ses composants sur les microtubules permettant ainsi d'augmenter ses chances d'être acquis par le vecteur (Martiniere *et al.*, 2013).

Il est tout de même important de préciser qu'à ce jour les études de transmission sont majoritairement effectuées sur des systèmes modèle en conditions de laboratoire (Alexander *et al.*, 2013).

2.1.2.2. Gamme d'hôte

- ***Généralistes vs. spécialistes***

La gamme d'hôte correspond à l'ensemble des espèces sensibles à un agent pathogène. Connaître cette gamme d'hôte est important pour comprendre et prédire l'impact des agents pathogènes. On distingue les agents pathogènes généralistes (qui vont infecter une large gamme d'espèces hôtes) des agents pathogènes spécialistes (qui vont infecter une gamme d'hôte étroite voire une seule espèce). Les virus ne peuvent maximiser leur fitness dans tous les hôtes car cela a un coût, on parle de trade-off adaptatif (Elena *et al.*, 2014) (Figure SB.13). Cependant, même s'il est généraliste, un virus n'aura pas la même fitness dans chacun de ses hôtes (Betancourt *et al.*, 2011; Elena

et al., 2014; Power *et al.*, 2011; Sacristan *et al.*, 2005). Par ailleurs, quand un saut d'hôte se produit, les mutations ayant permis l'adaptation virale au nouvel hôte peuvent aussi se traduire par une réduction de la fitness du virus dans l'hôte d'origine ; ce qui conduit à l'occurrence d'un trade-off adaptatif de la virulence. Les virus généralistes (parmi lesquels on compte le *Bean yellow mosaic virus*, le *Tomato spotted wilt virus* ou encore le *Turnip yellows virus*) sont supposés avoir évolué au sein de communautés de plantes sauvages diversifiées, alors que les virus spécialistes seraient plutôt issus de communautés présentant une faible diversité (Jones, 2009). Dans une étude de la prévalence de 5 phytovirus sur une gamme de 21 plantes hôtes, il a été démontré que les plantes hôtes étaient préférentiellement infectés par certains virus, et que le virus le plus sélectif (c'est-à-dire le plus spécialiste) s'est révélé être le plus prévalent dans l'ensemble des individus testés de sa gamme d'hôte restreinte, ce qui suggère que la spécialisation permettrait une adaptation optimale à l'hôte ciblé (Malpica *et al.*, 2006). Plusieurs hypothèses relatives à l'impact des virus généralistes/spécialistes sur les communautés de plantes ont été élaborées. Il a par exemple été proposé qu'un virus généraliste, qui par définition va infecter une large gamme d'hôtes, aura plus de chances d'avoir un impact important sur la composition des communautés de plantes (Alexander, 2010; Vincent *et al.*, 2014). Malheureusement les données disponibles sur ces virus restent fragmentaires car la majorité des travaux de recherche ont souvent été menées sur une seule espèce hôte ce qui rend difficile la généralisation des résultats (Power, 2008). Par ailleurs, au-delà du fait qu'il soit généraliste ou spécialiste, un phytovirus va avoir un effet différent sur la composition d'une communauté selon qu'il infecte une espèce rare ou une espèce dominante (Alexander, 2010).

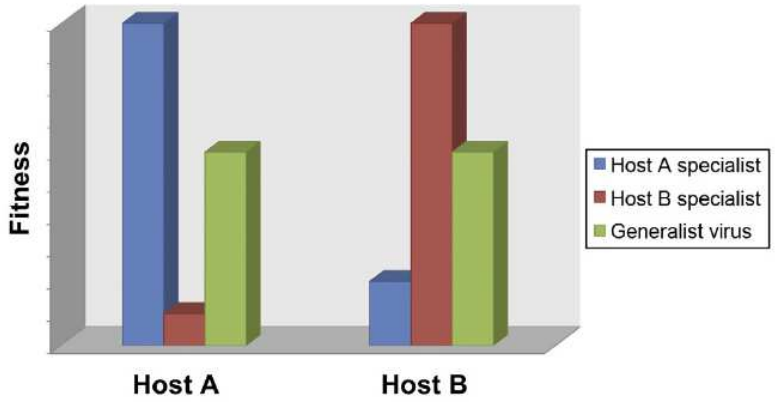


Figure SB.13 : Représentation du trade-off adaptatif. Sur cette figure sont représentées les fitness attendues pour des virus spécialistes et généralistes. Les virus spécialistes ont une fitness élevée dans leur hôte principal auquel ils sont adaptés, et une fitness diminuée dans l'hôte alternatif. Le virus généraliste a quant à lui une fitness intermédiaire dans les deux hôtes. D'après Elena *et al.*, 2014.

- ***Infections mixtes ou co-infection***

Plusieurs études ont montré que des virus différents partagent certaines espèces de leur gamme d'hôte (Hull, 2002; Martin and Elena, 2009; Roossinck, 2005; Szathmary, 1992). Ces virus peuvent alors infecter la même plante au même moment, on parle de co-infections ou d'infections mixtes. Les interactions entre des virus co-infectant la

même plante peuvent être théoriquement multiples. Admettons que les virus A et B co-infectent une plante X. Le virus A peut avoir un effet négatif sur la fitness du virus B (compétition) ou alors le virus A peut augmenter la fitness du virus B (synergisme via la complémentation ou la facilitation) (Power, 2008). Dans l'étude de la prévalence de 5 phytovirus sur une gamme de 21 plantes hôtes, il a par exemple été démontré que les virus co-infectant le même d'hôte se comportaient de manière synergique (Malpica *et al.*, 2006). Par ailleurs, il existe des cas où certains virus ne peuvent être transmis que sous la forme de complexes notamment par emprunt de la protéine de capside (Cooper and Jones, 2006; Lecoq, 1994; Querci *et al.*, 1997; Syller and Marczewski, 2001). Par conséquent, les infections mixtes peuvent avoir divers effets : une aggravation de l'accumulation virale et de la virulence, un élargissement de la gamme d'hôte, une augmentation du taux de transmission, une amélioration du tropisme cellulaire (capacité à infecter certains types de cellules), etc. (Elena *et al.*, 2014).

- ***Rôle de la vection sur la gamme d'hôte***

Le nombre d'espèces vectrices, leur type (nématode, champignon, arthropode), et leur gamme d'hôte sont des déterminants primordiaux pour la gamme d'hôte du virus (Power, 2008; Power and Flecker, 2003). Bien que les virus de plantes soient souvent généralistes en terme de gamme d'hôte, il seraient plutôt spécialistes en terme de gamme de vecteurs (Power, 2008; Power and Flecker, 2003) (Figure SB.14). Parmi les virus de plantes transmis par vecteur, 60% sont transmis par une seule espèce d'insecte alors que moins de 10% des phytovirus n'infectent qu'une seule espèce de plante (Power, 2008; Power and Flecker, 2003). La gamme d'hôte des virus est tributaire de celle de leurs vecteurs (Elena *et al.*, 2014). Ainsi, toute expansion de la gamme d'hôte du vecteur peut conduire à une expansion de la gamme d'hôte des virus qu'ils transmettent (Goldbach and Peters, 1994; Harrison and Robinson, 1999) et à l'occurrence de sauts d'hôtes (Elena *et al.*, 2014).

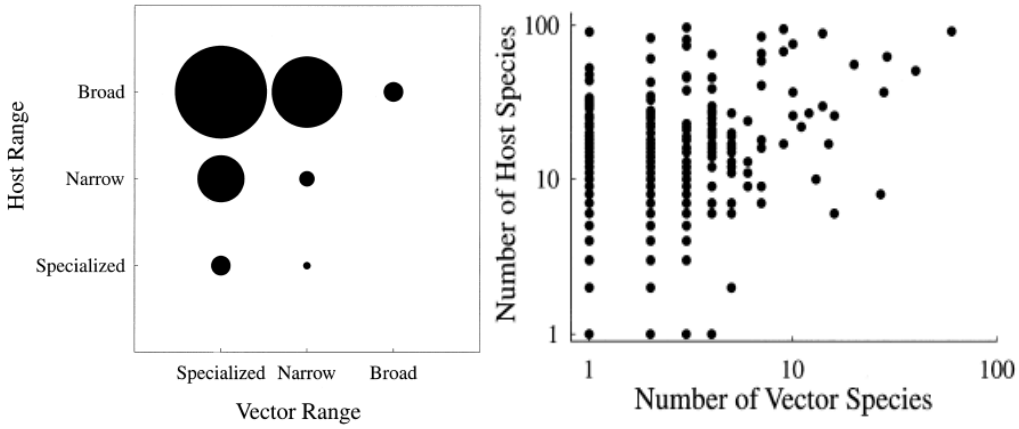


Figure SB.14 : Nombre d'hôtes et nombre de vecteurs des phytovirus transmis par vecteurs (n=474). a) Relation entre gamme d'hôte et gamme de vecteur des phytovirus. b) Nombre d'espèces hôtes en fonction du nombre d'espèces vectrices pour chaque virus. Chaque point représente une espèce virale. D'après Power and Flecker, 2003.

2.2. Les interactions plantes/virus à l'échelle des communautés et du paysage

Depuis 40 ans les études sur le rôle que jouent les virus dans la structure et l'évolution des communautés de plantes n'ont cessé d'augmenter (Burdon and Thrall, 2013).

Les virus ont la capacité de changer la composition des communautés de plantes (hôtes et non-hôtes) en influant notamment sur la densité des espèces hôtes directement ou indirectement. En retour, les communautés de plantes vont elles aussi avoir un impact sur les populations virales (Alexander, 1998; Alexander *et al.*, 2013; Borer *et al.*, 2007; Corbin and D'Antonio, 2004; Dunn *et al.*, 2012; Malmstrom *et al.*, 2005a; Malmstrom *et al.*, 2005b; Power and Mitchell, 2004). Ainsi, les interactions plante-virus ne doivent pas être seulement considérées au niveau de l'individu ou de la communauté, mais elles doivent l'être aussi à l'échelle du paysage en tenant compte de tous les facteurs (biotiques et abiotiques) qui peuvent les influencer.

2.2.1. Rôle de la biodiversité et de la densité des plantes sur les dynamiques virales

La densité des espèces hôtes peut affecter directement l'incidence et la sévérité d'une maladie via des effets sur le taux de rencontre hôte-agent pathogène (Burdon and Chilvers, 1982). En effet les variations de densité induisent des changements du nombre d'hôtes disponibles dans l'espace et dans le temps et de la distance entre hôtes sensibles (Gilbert, 2002; Keesing *et al.*, 2010; Keesing *et al.*, 2006).

Par ailleurs, deux hypothèses antagonistes lient la biodiversité des espèces hôtes aux dynamiques virales et aux émergences (Keesing *et al.*, 2006) :

- **L'effet de dilution** : l'augmentation de la diversité des espèces dans la communauté de plantes serait corrélée de façon négative au risque d'épidémie. Dit autrement, une diminution de la diversité peut amener à une abondance accrue de l'espèce hôte focale ce qui facilite la transmission de l'agent pathogène. Par exemple, une étude sur le piment sauvage au Mexique a démontré que la prévalence des virus qui lui sont associés augmente avec les niveaux d'anthropisation (qui se traduit par une diminution de la diversité), ce qui laisse suggérer une corrélation négative entre la biodiversité et le risque d'épidémie (Pagan *et al.*, 2012; Rodelo-Urrego *et al.*, 2013). Cet effet de dilution est l'hypothèse majoritairement invoquée dans les études d'influence de la biodiversité des hôtes sur les dynamiques des agents pathogènes (Keesing *et al.*, 2006).

- **L'effet d'amplification** : l'augmentation de la diversité via un accroissement des réservoirs potentiels d'agents pathogènes serait corrélée positivement au risque d'épidémie. Par exemple, les *Barley yellow dwarf virus* et *Cereal yellow dwarf virus* (B/CYDV) qui sont des virus ayant une large gamme d'hôte au sein des graminées ont

souvent été décrits comme étant soumis à l'effet d'amplification bien que dans quelques études l'effet de dilution soit aussi démontré (reporté dans la revue de Elena *et al.*, 2014).

Ainsi, la validation de ces hypothèses va dépendre de la gamme d'hôte de l'agent pathogène. En effet, si l'augmentation de la diversité se traduit par une augmentation du nombre d'espèces hôtes, alors le risque d'épidémie est augmenté. *A contrario*, si l'augmentation de la diversité se traduit par l'augmentation du nombre d'espèces non-hôtes, alors le risque d'épidémie sera diminué (Keesing *et al.*, 2006).

Il est clair que les pratiques agricoles ont mené à une perte de diversité génétique et à l'augmentation de la densité des espèces cultivées (ceci est appelé la simplification écologique), on peut alors faire l'hypothèse que cela a accru la vulnérabilité des cultures aux agents pathogènes par l'augmentation de la taille de leur inoculum et un accroissement de leur adaptation à l'hôte cultivée (Edwards, 1996; Elena *et al.*, 2014).

Alicia Keesing argumente le fait que l'effet de dilution est plus probable (i) lorsque la transmission de l'agent pathogène est fréquence-dépendante (c'est-à-dire qu'elle dépend de la proportion des individus hôtes par rapport aux individus non-hôtes et non de la densité absolue des individus hôtes), (ii) quand la transmission intraspécifique est plus importante que la transmission interspécifique, et surtout (iii) quand l'hôte le plus compétant est l'hôte le plus répandu (Keesing *et al.*, 2006). Il faut également tenir compte de l'impact de la diversité végétale sur les populations de vecteurs, en effet, la perte de diversité des plantes peut amener à des changements dans l'abondance, le comportement et la biologie des vecteurs (Keesing *et al.*, 2010). Ainsi, avant d'invoquer l'une des deux hypothèses, il est important de bien connaître les modalités de la transmission de l'agent pathogène et de sa gamme d'hôte.

2.2.2. Impact des phytovirus sur la compétition entre plantes

La définition de la compétition sur laquelle nous nous appuyons dans ce manuscrit est la suivante: « Interaction quelconque entre individus qui amène à une réduction de la fitness d'un ou plusieurs de ces individus » (Alexander, 1998). Les interactions compétitives chez les plantes varient en fonction de facteurs biotiques et abiotiques. Des différences de sensibilité ou de tolérance à un agent pathogène peuvent par exemple induire des interactions compétitives différentes selon que les plantes soient infectées ou non (Gilbert, 2002). Les agents pathogènes peuvent ainsi avoir un effet sur les interactions intraspécifiques et interspécifiques chez les plantes, aboutissant *in fine* à un changement dans la composition des communautés végétales (Alexander, 1998).

Malheureusement, la documentation sur ce sujet concernant les virus de plantes est restreinte et la majorité des études disponibles concernent les champignons phytopathogènes (Alexander, 1998). Cependant, des études théoriques permettent d'examiner les causes et les effets potentiels d'une telle compétition. Au niveau

intraspécifique, les plantes non infectées vont être en compétition avec celles infectées et leur taux de croissance et leur reproduction vont être accrus ; Cornelis T. De Wit appelle cela « l'effet compensatoire » car il y a compensation des pertes numériques engendrées par la maladie par les plantes saines. Cornelis T. De Wit suggère alors qu'il existe une corrélation entre la variation de la résistance et la capacité compétitive (De Wit, 1986). En effet, si une plante sensible A domine la communauté et une plante B résistante est minoritaire (car on suppose qu'il y a un coût à la résistance), quand la population A est touchée par un pathogène, la population B va devenir majoritaire, il y a alors un déplacement (switch) de capacité compétitive dû à la présence de la maladie (Burdon and Chilvers, 1982). Si l'on est dans le cas de deux population sensibles, Robert D. Holt fait l'hypothèse que l'accroissement de densité d'une des deux populations va permettre d'accroître la densité du parasite qui va alors avoir un effet négatif sur la croissance de la deuxième population hôte, on appelle cela la « compétition apparente » (Alexander, 1998; Holt *et al.*, 1994; Holt and Pickering, 1985; Mitchell and Power, 2006). Bien évidemment, la nature des interactions compétitives entre plantes en présence d'une maladie dépend des patrons spatiaux de l'incidence de la maladie et d'une distance seuil à partir de laquelle on peut considérer que les plantes ne sont plus en compétition (Alexander, 1998). Les plantes introduites et les agents pathogènes qui les affectent vont intervenir dans cette compétition (cf. section 2.2.3.1).

2.2.3. Rôle de l'Homme sur les dynamique virales

Plusieurs travaux récents ont suggéré l'influence grandissante de l'Homme sur la détérioration des espaces naturels (Ellis *et al.*, 2010; Ellis and Ramankutty, 2008; Kareiva *et al.*, 2007). De ce fait, il est communément considéré que les activités humaines ont une influence sur les équilibres biologiques de ces espaces conduisant probablement dans certains cas à l'émergence de maladies virales.

L'intensification des cultures couplée à l'irrigation amène à l'expansion des populations de vecteurs et à une interaction positive entre ces vecteurs et les populations de virus (Anderson *et al.*, 2004). La diversification se traduit par l'introduction de nouveaux hôtes, et la globalisation amène à l'introduction de nouveaux pathogènes et de vecteurs, ce qui permet la propagation de maladies exotiques. Les activités humaines peuvent donc avoir des effets directs mais aussi des effets indirects via par exemple le changement climatique (Figure SB.15) (Anderson *et al.*, 2004).

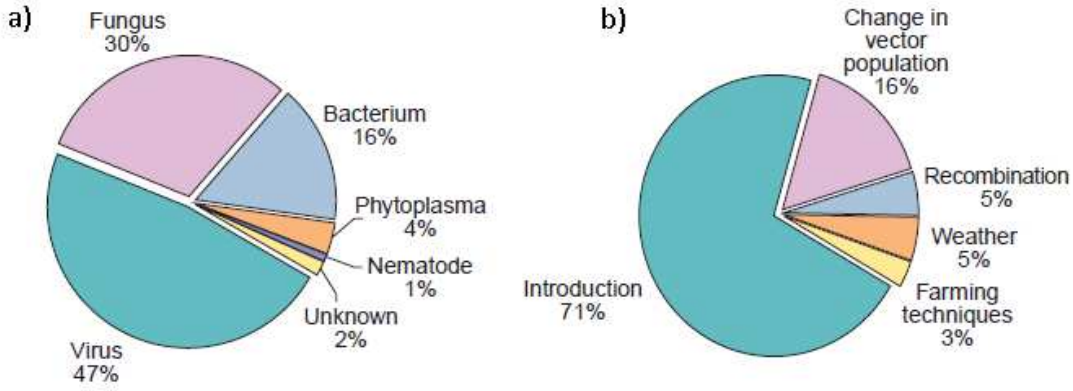


Figure SB.15: Caractéristiques des agents pathogènes émergents et des facteurs impliqués dans l'émergence des maladies des plantes. a)Proportion des organismes à l'origine de maladies émergentes de plantes. b) Facteurs responsables des maladies virales émergentes de plantes. D'après Anderson *et al.*, 2004.

2.2.3.1. Rôle des plantes introduites et des invasives sur les émergences

Les communautés de plantes ont fortement été modifiées par l'action de l'Homme depuis sa sédentarisation. L'introduction de plantes exotiques dans une aire géographique différente de celles de leur origine est un des facteurs ayant entrainé ces modifications. Une plante vivant dans son aire d'origine est appelée « native » ou « indigène ». Il ne faut pas confondre les termes « exotique » et « invasive », en effet, une plante exotique ne sera pas forcément invasive dans l'aire d'introduction, et une plante native peut très bien être invasive.

Les espèces indigènes sont supposées être particulièrement vulnérables aux agents pathogènes introduits avec lesquels elles n'ont pas coévolué. Les introductions de nouvelles espèces de plantes hôtes amènent souvent à l'émergence de nouveaux virus (Anderson *et al.,* 2004; Cooper and Jones, 2006; Fargette *et al.*, 2006; Jones, 2009; Webster *et al.*, 2007). A titre d'exemple, le *Bean yellow mosaic virus* (BYMV) a probablement été introduit en Australie via des bulbes de glaïeul et infecte désormais une large gamme de légumineuses indigènes dans la région floristique du Sud-Ouest mais aussi dans l'Est de l'Australie ; parmi elles, *Kennedia prostrata* est sévèrement touchée (Mckirdy *et al.*, 1994; Webster *et al.*, 2007).

Plusieurs travaux théoriques et expérimentaux ont été conduits pour mieux comprendre les mécanismes associés à l'émergence de maladies virales en relation avec l'introduction de plantes exotiques (Figure SB.16) (Jones, 2009). En écologie phytovirale, une des hypothèses les plus couramment évoquée et la plus souvent démontrée est celle de l' « Enemy release » qui statue que lorsqu'une plante exotique est introduite dans une nouvelle aire elle n'est plus en interaction avec ses ennemis naturels (agents pathogènes, prédateurs…) présents dans son aire native. La diminution des pressions sélectives peut se traduire par une aptitude accrue à devenir invasive dans la communauté où elle a été introduite (Mitchell and Power, 2003). A l'inverse de l' « Enemy release », la théorie de « Biotic resistance » suppose que les interactions de

l'espèce introduite avec les espèces natives, incluant les ennemis naturels (pathogènes indigènes, herbivores…) limitent son potentiel à envahir la communauté (Mitchell and Power, 2003).

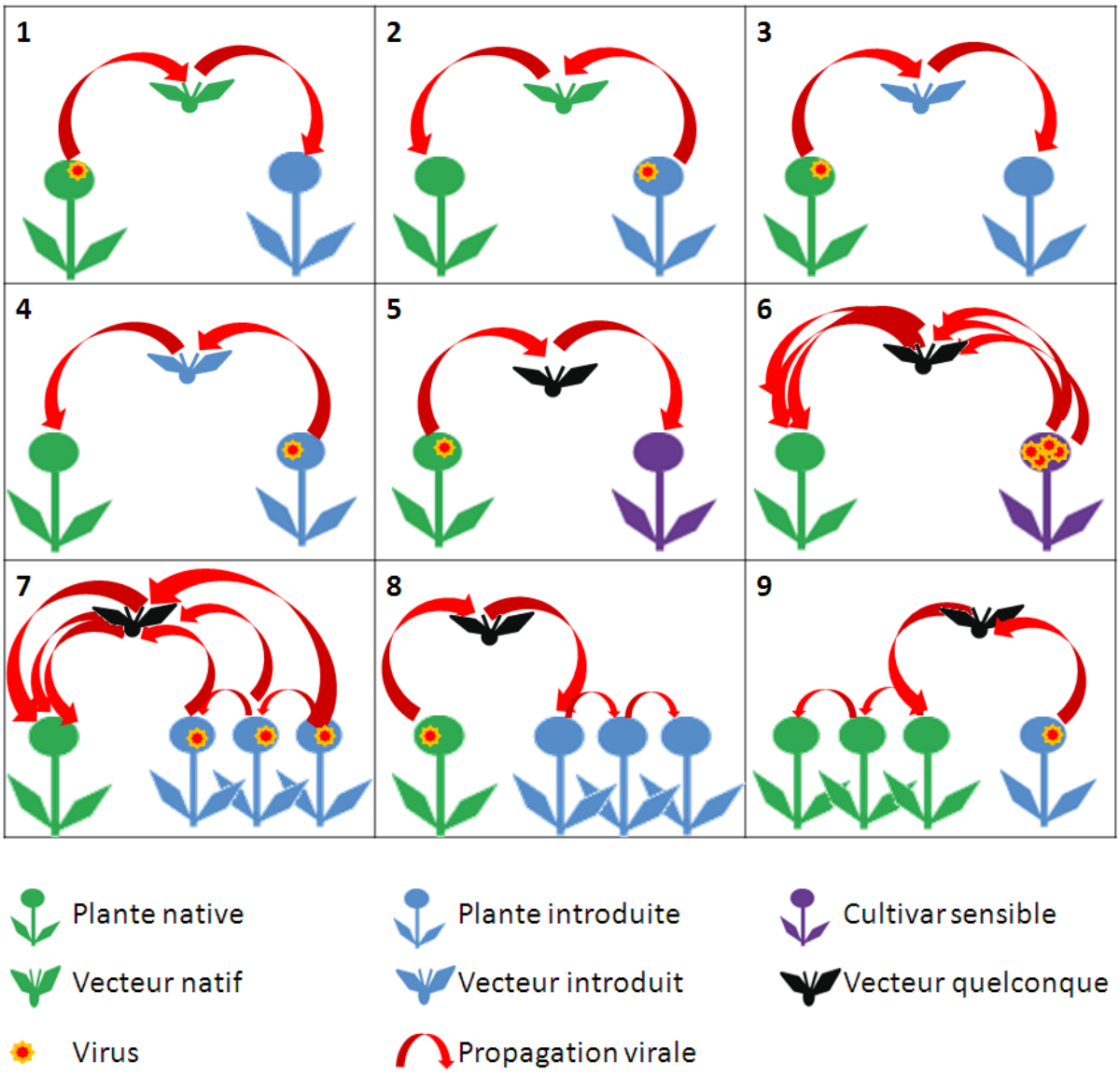


Figure SB.16: Schématisation des scenarios d'émergence entre plantes natives et introduites. Le scénario 6 correspond au phénomène de pathogen spillover. Les 7 et 8 correspondent à l'intensification/extension et diversification agricole, dans le cas du 7, l'augmentation de réservoirs agricoles est augmenté, dans le cas du 8, les pratiques agricoles réduisent la distance entre le réservoir natif et les cultures, le scénario 9 correspond à une réduction des distances entre plante introduite et natives ce qui amène à une propagation dans les natives. Inspiré par la revue de Jones, 2009.

Une autre théorie, dite du « Pathogen spillover », statue qu'un virus s'étant fortement accumulé dans une plante réservoir peut ensuite se répandre sur un hôte moins approprié pour sa multiplication mais beaucoup plus sensible, ceci se traduisant alors par l'occurrence de dégâts importants sur ce second hôte (Daszak *et al.*, 2000; Power and Mitchell, 2004). Cette théorie a été vérifié expérimentalement grâce au « pathosystème » B/CYDV / graminées de Californie (Borer *et al.*, 2007; Malmstrom *et al.*, 2005b). Ces études ont démontré que l'introduction d'*Avena fatua* en Californie,

espèce tolérante au B/CYDV, s'est traduite par un accroissement de la prévalence du virus généraliste BYDV-PAV au sein des espèces natives de graminées (Power and Mitchell, 2004). L'espèce *Avena fatu*a s'est révélée être à la fois un très bon réservoir présentant de forts titres viraux et un hôte très attractif pour les vecteurs. Ces deux caractéristiques ont conduit à une forte augmentation des accumulations du virus au sein de l'écosystème californien (Lowry, 2007; Malmstrom *et al.*, 2005b). L'incidence de la maladie dans *Elymus glaucus* (espèce pérenne native) a ainsi augmenté ce qui a mené à un remplacement des graminées pérennes natives par l'espèce annuelle introduite dans la communauté (Malmstrom *et al.*, 2005b) et ceci même si ces annuelles sont considérées comme étant de faibles compétitrices en l'absence du virus (Borer *et al.*, 2007; Corbin and D'Antonio, 2004; Seabloom *et al.*, 2003).

Bien que l'essentiel des études sur les interactions natives/introduites ait été effectué sur le B/CYDV, des travaux expérimentaux basé sur l'étude de l'effet d'autres virus exotiques sur des plantes natives commencent à apparaître dans la littérature. A titre d'exemple, une étude récente s'est intéressée à 15 espèces de plantes natives australiennes (réparties dans 8 familles) inoculées par 6 virus généralistes introduits. Il s'est avéré de manière générale que ces virus ont une gamme d'hôte très large (par exemple le *Cucumber mosaïc virus* a développé une infection au niveau de 7 des 8 familles testées). De plus la majorité des virus testés ont induit des symptômes pouvant être très sévères et ont conduit à une diminution de la biomasse et de la croissance de certaines plantes inoculées (Vincent *et al.*, 2014).

2.2.3.2. Rôle des vecteurs introduits sur les émergences

Un agent pathogène peut être introduit dans un nouvel écosystème sans pour autant causer une émergence. Un des facteurs principal amenant aux émergences virales est l'introduction de vecteurs. Par exemple, le *Citrus tristeza virus* (CTV) a été introduit en Amérique du Sud entre 1927 et 1930, mais c'est l'introduction de son vecteur *Toxoptera citricidus* provenant d'Asie qui a mené à son émergence (Bar-Joseph *et al.*, 1979). Un cas souvent étudié dans le cadre de l'influence d'un vecteur sur l'émergence de maladies est celui du vecteur des Begomovirus, *Bemisia tabaci*. En effet, le biotype-B étant plus fécond que les autres biotypes, il a conduit à la propagation à longue distance et mondiale du *Tomato yellow leaf curl virus* qui a fait des dommages considérables sur les cultures de tomate (Figure SB.17) (Perefarres *et al.*, 2012).

Dans un autre contexte, en 1970, l'expansion de la culture de *Glycine max* (soja) au Brésil et en Argentine a conduit à l'émergence d'agents pathogènes touchant le haricot *Phaseolus vulgaris* : le *Bean golden mosaic virus* a mené à des pertes de rendement de 85% dans le sud du Brésil et le *Bean dwarf mosaic virus* à l'effondrement de la culture du haricot en Argentine. En effet, le vecteur de ses 2 pathogènes, *Bemisia tabaci* se reproduit sur les plants de soja, l'intensification des cultures de soja a alors permis d'accroitre les populations de vecteurs, ce qui a *in fine* conduit à l'émergence des deux maladies (Costa, 1975; Morales and Anderson, 2001).

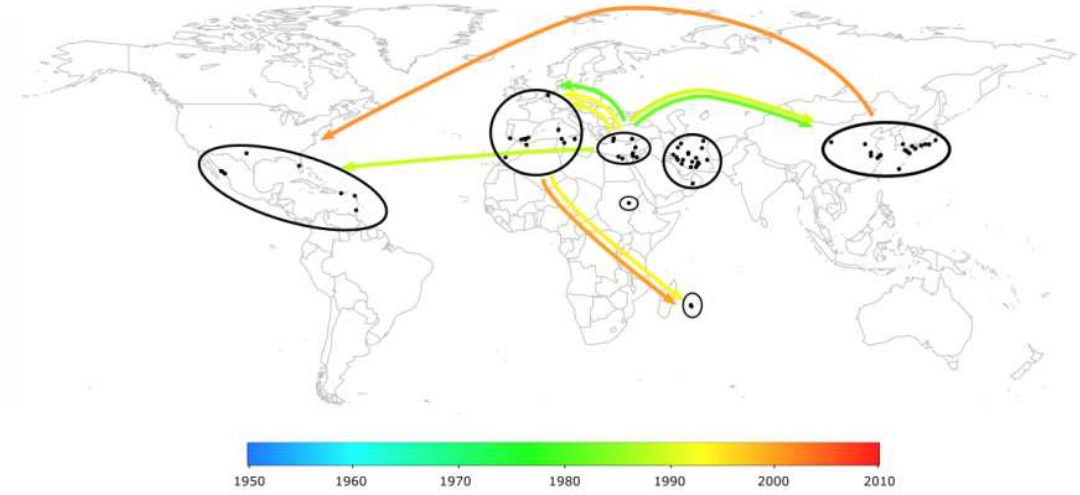


Figure SB.17: Propagation à travers le monde du *Tomato yellow leaf curl virus* (TYLCV) inférée via des analyses de phylogéopgraphie de la protéine de capside et des génomes entiers. Selon l'échelle au dessous de la figure, la couleur des flèches correspond à la date approximative des évènements d'introduction. D'après Perefarres *et al.*, 2012.

2.2.3.3. Rôle du pâturage sur les dynamiques virales

Les activités humaines amènent à une redistribution mondiale des prédateurs et des herbivores (qu'elle soit délibérée ou accidentelle) (Mack *et al.*, 2000). Or, les herbivores sont connus pour leur effet sur les communautés végétales (Keesing *et al.*, 2006; Mesleard *et al.*, 2011). Il a été par ailleurs spéculé que le pâturage peut avoir un impact sur les dynamiques virales. Toutefois les études à ce sujet sont peu nombreuses (Borer *et al.*, 2009; Malmstrom *et al.*, 2006; Power *et al.*, 2011). En théorie, si on fait abstraction des préférences alimentaires des prédateurs, une prédation directe sur une population hôte est sensée réduire les populations d'agents pathogènes qui lui sont associés (McCallum *et al.*, 2001; Ostfeld and Holt, 2004; Packer *et al.*, 2003). Une étude testant l'effet à long terme de l'exclusion des herbivores vertébrés sur la prévalence du B/CYDV a infirmé cette théorie. En effet, la prévalence du B/CYDV est 4 fois plus importante en présence du pâturage que dans les aires où il est absent. Il a ainsi été démontré que les herbivores vertébrés ont une influence indirecte sur la prévalence virale. En effet, les herbivores en question ont une préférence pour les plantes pérennes de la communauté végétale qui sont en l'occurrence de mauvais hôtes ou des plantes non-hôtes. La diminution de l'abondance des plantes pérennes va permettre aux annuelles, des hôtes très compétents, d'envahir la communauté. Ainsi, les herbivores accroissent indirectement le risque d'infection dans la communauté végétale (Borer *et al.*, 2009). Une autre étude a testé l'effet potentiel du pâturage sur deux espèces de graminées (*Nassella pulchra* et *Elymus multisetus*) ayant des sensibilités différentes à l'infection par le B/CYDV ; *E. multisetus* est plus sensible et moins abondante. L'effet d'un pâturage non-sélectif a été testé via la coupe des plantes par des cisailles et ce sur des communautés de plantes soumises ou non à la compétition via l'introduction d'une espèce dominante *Bromus hordaceus*. Ces travaux ont démontré que le pâturage atténue

les effets de la compétition permettant ainsi d'accroître la survie des graminées testées. En outre, dans les communautés ou l'abondance de l'espèce compétitrice est faible, le pâturage permet d'accroître la survie des graminées virosées alors qu'intuitivement on aurait pu supposer que les effets combinés de l'infection et du pâturage limiteraient la survie de ces graminées. Ces résultats ont ainsi montré que sous certaines conditions le pâturage non-sélectif peut améliorer la survie des plantes infectées, ce qui a un impact sur la persistance du virus dans la communauté de graminées (Malmstrom *et al.,* 2006). Par ailleurs, une autre étude menée sur les effets du pâturage a démontré que les herbivores pouvaient intervenir dans la transmission de certains virus via la consommation et la mastication (cas de la transmission du *White clover mosaic virus* à *Trifolium repens*) ou via le piétinement (cas de la transmission du *Suterraneum mottle virus* à *Trifolium subterraneum*) (Mckirdy *et al.*, 1994).

2.2.3.4. Rôle de la domestication des plantes sur les dynamiques virales

La domestication des plantes par l'Homme pour les besoins de l'agriculture a engendré de fortes pressions de sélection sur celles-ci les conduisant à être plus productives, peu diversifiées, mais aussi dépendantes de l'Homme pour leur survie (Jones, 2009). Ainsi, la théorie de l'allocation prédit que l'accroissement des taux de croissance des plantes a un coût qui va conduire à une diminution de l'allocation des ressources dans les défenses (Herms and Mattson, 1992). Selon cette théorie, la sélection de certains traits tels que la productivité peut compromettre les défenses de la plante (Kennedy, 1992; Schrotenboer *et al.*, 2011). Par exemple, il a été démontré que des cultivars de *Panicum virgatum* sélectionnés pour la production de biocarburants à partir de plants sauvages étaient à la fois plus sensibles aux infections de B/CYDV et plus attractifs pour les pucerons vecteurs que les plants sauvages parentaux (Schrotenboer *et al.*, 2011).

2.2.3.5. Impact des changements climatiques sur les dynamiques phytovirales

Plusieurs études ont montré que le climat peut influencer directement et indirectement les interactions plantes-agents pathogènes et avoir un effet sur l'incidence et les distributions spatio-temporelles des maladies de plantes que ce soit dans les écosystèmes naturels ou anthropisés (Stern and Taylor, 2007). Les changements climatiques peuvent conduire à l'émergence de maladies au travers de changements graduels du climat, en altérant par exemple la distribution des vecteurs ou en augmentant les stress hydriques ou thermiques des plantes, ou de changements brusques avec l'occurrence d'évènements climatiques extrêmes (Anderson *et al.*, 2004).

Les changements climatiques vont agir de manière directe sur la physiologie des plantes individuelles et des communautés en induisant des stress abiotiques (Sutherst *et al.*, 2011). Les dynamiques des agents pathogènes peuvent alors être modifiées dans le sens où leur mortalité en hiver peut être réduite, le nombre de générations par an peut être augmenté, et des sauts d'hôtes et changements d'aires géographiques peuvent être

plus fréquents (Alexander, 2010). Sous l'influence du changement climatique, ces pathogènes vont donc évoluer de manière accélérée via des temps d'incubation plus courts (Sutherst *et al.*, 2011). La majorité des études sur l'effet du changement climatique sur les interactions hôtes-agents pathogènes ont été faites sur champignons phytopathogènes pour lesquels il a été démontré que la température, les nutriments et le CO_2 avaient un impact fort sur leurs dynamiques (Coakley *et al.*, 1999; Garrett *et al.*, 2006; Harvell *et al.*, 2002). On peut tout de même citer deux études qui ont montré que des conditions de sécheresse augmentaient les effets délétères du *Maize dwarf mosaic virus* (Olson *et al.*, 1990) et du *Beet yellows virus* (Clover *et al.*, 1999). Par ailleurs, une étude récente a permis de démontrer que les titres viraux du BYDV dans le blé augmentaient avec un accroissement de la température (Nancarrow *et al.*, 2014). Il a notamment été suggéré que le changement climatique pouvait affecter la résistance des plantes aux agents pathogènes (Garrett *et al.*, 2006). Par exemple, des taux en CO_2 et O_3 élevés affecteraient l'efficacité des résistances hôtes et supprimeraient l'induction de la résistance par les agents pathogènes (Garrett *et al.*, 2006).

En plus d'avoir un effet sur la physiologie des plantes hôtes, le climat, via les fluctuations de températures, influence grandement le développement et la distribution des populations de vecteurs, et par conséquent un changement dans les populations de vecteurs peut mener à un changement dans les populations virales (Anderson *et al.*, 2004).

En retour, les maladies de plantes peuvent être utilisées comme indicateurs du changement climatique (Garrett, 2009). Toutefois, peu d'études ont été effectuées à ce jour sur ce thème (Garrett *et al.*, 2006). Enfin, il a récemment été considéré que les effets du changement climatique seraient globalement moins importants que l'impact des pratiques agricoles, et seraient différents selon les aires d'étude, empêchant une généralisation simple de la quantification de ces phénomènes perturbateurs (Coakley *et al.*, 1999).

2.3. Les virus dans le milieu sauvage et l'interface agro-écologique

2.3.1. Définition du milieu sauvage et du milieu cultivé

Tout d'abord il est important de définir ce qu'est le milieu sauvage. Helen M. Alexander suggère que la première distinction entre le milieu sauvage et le milieu agricole est que la dynamique des populations de plantes du milieu sauvage et leurs trajectoires évolutives ne sont pas soumises au contrôle direct de l'Homme (Alexander *et al.*, 2013). En effet dans le milieu cultivé, les plantes sont « contrôlées » à divers degrés par l'Homme : travail du sol, sélection des variétés, densité des graines, rotations culturales, apport d'engrais, de pesticides, etc. Dans le milieu sauvage, les graines qui vont se développer et persister dans les générations futures ne sont soumises qu'aux phénomènes de migration, dérive et sélection naturelle. De ce fait, le peuplement du

milieu sauvage est plus diversifié, complexe et hétérogène que le peuplement du milieu agricole. Les plantes qui se côtoient dans le milieu sauvage peuvent avoir des âges et des stades de développement différents, et leur répartition spatiale et leur densité n'est pas régulière. En effet, contrairement aux espèces cultivées, les populations naturelles de plantes ont des arrangements spatiaux complexes avec de 10 à 100 espèces (natives et exotiques) qui sont génétiquement très diverses, elles présentent également une densité d'individus plus faible (Alexander, 2010; Alexander *et al.*, 2013; Burdon and Thrall, 2008). Le milieu sauvage subit une multitude de stress : compétition entre plantes, herbivorie, environnement biotique et abiotique non contrôlé par l'Homme (Alexander *et al.*, 2013). Les communautés agricoles quant à elles présentent une faible diversité, de fortes densités, et sont génétiquement contrôlées par l'Homme (Burdon and Thrall, 2008).

2.3.2. Les virus dans le milieu sauvage

2.3.2.1. Pourquoi s'intéresser aux phytovirus du milieu sauvage ?

L'étude des phytovirus du milieu sauvage a souvent été négligée. Ceci est en partie dû à l'absence de symptômes apparents sur beaucoup de plantes sauvages (Prendeville *et al.*, 2012; Remold, 2002; Roossinck, 2010; Roossinck, 2012b) et au fait que les écologistes sont peu familiers aux techniques de détection des virus au laboratoire (Malmstrom *et al.*, 2011). Mais la raison principale est qu'une attention plus grande a été portée aux plantes d'intérêt agronomique (Cooper and Jones, 2006). Par conséquent, notre connaissance sur les interactions plantes-virus vient surtout des virus qui infectent une faible proportion de plantes, qui sont prévalents dans des cultures peu diversifiées et qui induisent des phénotypes marqués (symptômes, baisse des rendements) (Roossinck *et al.*, 2010; Wren *et al.*, 2006). Ce type de connaissance nous renseigne sur les interactions à court terme (quelques centaines à quelques milliers d'années) entre les virus et leurs hôtes mais ne nous permet pas de comprendre les interactions à long terme (centaines de milliers à millions d'années) entre plantes et virus (Roossinck *et al.*, 2010; Wren *et al.*, 2006). Il est fort probable que l'étude des virus du milieu sauvage pourrait éclairer ce pan de l'évolution virale qui reste encore inconnu.

Ce n'est que depuis les années 1970 que l'on a commencé à porter un intérêt aux virus infectant les plantes sauvages et les mauvaises herbes parce que l'on a supposé qu'ils partagent des vecteurs avec le milieu cultivé. On s'est alors intéressé aux mouvements viraux du milieu sauvage vers le milieu cultivé (Alexander *et al.*, 2013). Quelques études ont montré par ailleurs que des virus provenant du milieu cultivé provoquent des maladies au niveau des plantes sauvages (Webster *et al.*, 2007; Wylie *et al.*, 2013). Mais, globalement, très peu d'attention a été portée aux dynamiques virales intrinsèques au milieu sauvage. Une étude récente concernant 5 virus infectant 21 populations sauvages de *Cucurbita pepo* a démontré que 80% des infections ne causaient pas de symptômes apparents (Prendeville *et al.*, 2012). Cependant, bien que les symptômes ne soient pas toujours visuellement apparents sur les parties aériennes

de la plante, d'autres études ont lié l'infection virale à une fertilité réduite, et à une altération de la production de graines (Cooper and Jones, 2006).

2.3.2.2. Prévalence des virus dans le milieu sauvage

À ce jour, seulement deux études publiées ont fait un inventaire exhaustif de la diversité des virus dans le milieu sauvage. De façon inattendue, ces publications révèlent que près de la moitié des plantes sauvages analysées contenaient des virus, ou du moins des reads / contigs viraux (Muthukumar *et al.*, 2009; Roossinck *et al.*, 2010). Pourtant les épidémies en milieu sauvage sont rares (Cooper and Jones, 2006). Les facteurs qui expliquent la rareté des épidémies dans le milieu sauvage sont : (i) un effet de dilution des plantes sources (en fonction de leur distribution) dans les communautés et donc une séparation géographique entre plante source et hôte potentiel, (ii) une diminution des populations de vecteurs due à la présence de prédateurs et de parasites naturels, et (iii) la présence de divers génotypes pour une même espèce de plante avec un panel plus important de gènes de résistance (Cooper and Jones, 2006). Toutefois, la fréquence des infections virales en milieu sauvage est plus difficile à déterminer que dans les cultures, car les stades de croissance des plantes sauvages sont hétérogènes; il se peut que les différents stades ne soient pas affectés de la même manière par l'infection virale. De plus, l'exposition à la source d'inoculum dépend de la position géographique et des conditions environnementales qui sont instables (Cooper and Jones, 2006). Il n'y a par ailleurs aucune surveillance structurée de routine et standardisée des communautés de plantes sauvages (Cooper and Jones, 2006).

2.3.2.3. Les plantes sauvages : des réservoirs viraux par excellence ?

Les plantes non-cultivées ont depuis longtemps été suspectées d'être à l'origine de maladies émergentes en étant des protagonistes-clé dans les phénomènes de sauts d'hôtes (réservoir initiaux) mais aussi des réservoirs viraux alternatifs lors des phénomènes de résurgence des maladies (Anderson *et al.*, 2004). A titre d'exemple, le *Maïze streak virus* a probablement émergé à partir de plantes sauvages après l'introduction du maïs en Afrique (Monjane *et al.*, 2011). Un autre exemple de plante sauvage jouant le rôle de réservoir viral alternatif est celui de *Sonchus oleraceux* en Australie ; son éradication à 150 m des champs de laitue a fait décroitre la prévalence du *Lettuce necrotic yellows virus* de 75% à 6% (Stubbs and Grogan, 1963). Comme nous l'avons décrit dans les chapitres précédents, les plantes sauvages peuvent également jouer le rôle « d'hôte spillover », ce qui fait également d'elles des réservoirs « d'accumulation » des virus (Power and Mitchell, 2004).

2.3.3. L'agro-écosystème : une interface dynamique

2.3.3.1. Définition et dynamique des virus à l'interface

À ce jour il n'existe plus d'écosystèmes vierges de perturbations anthropiques. Presque toutes les populations de plantes ont été influencées directement ou indirectement par les activités humaines (Ellis *et al.*, 2010; Ellis and Ramankutty, 2008;

Kareiva *et al.*, 2007) et notamment par l'intensification, la diversification et la globalisation de l'agriculture. Aujourd'hui, près de 39% des terres arables soit 4% des terres totales sont faites de monocultures de blé, de maïs et de riz (FAO, 2008).

De nombreux écosystèmes naturels se sont donc retrouvés adjacents à des espaces agricoles. Les bordures entre paysage agricole et naturel ont été appelées « interface agro-écologique » ou « agro-écosystème » (Burdon and Thrall, 2008) ou « lieu de rencontre » (encounters) (Jones, 2009). Au niveau de ces interfaces on retrouve donc des cultures, des plantes non-cultivées associées aux cultures (appelées adventices) et des plantes non-cultivées bordant les cultures qui peuvent être d'origine native ou exotique. Il est important de préciser que les adventices font partie intégrante du milieu cultivé, bien que les agriculteurs les considèrent comme sauvages ; elles sont adaptées à l'environnement agricole. La diversité végétale augmente selon un gradient du milieu cultivé vers le milieu sauvage (Malmstrom *et al.*, 2011). Cette interface est soumise à des facteurs biotiques et abiotiques qui se manifestent dans un continuum : dans les cultures, le milieu est riche en nutriments et en eau, plus on s'éloigne du milieu cultivé, plus les ressources sont limitantes ce qui amène donc à des phénomènes de compétition dans le milieu sauvage. Dans cette interface peuvent avoir lieu des échanges génétiques entre plantes cultivées et sauvages (Ellstrand, 2003) et des mouvements animaux (Alexander *et al.*, 2013). A plus large échelle, ces interfaces dynamiques sont influencées par les changements climatiques et sont des lieux privilégiés pour des phénomènes de redistribution de plantes, de virus, et de leurs vecteurs.

Les phytovirus sont des composants importants de cette interface, ils peuvent y être dispersés par les arthropodes pouvant ainsi coloniser indistinctement plantes sauvages et cultivées (Alexander *et al.*, 2013). En effet, les adventices et les plantes non-cultivées peuvent jouer le rôle de réservoirs de pathogènes qui influencent les épidémies des cultures. Par conséquent, dans le cas des cultures annuelles qui ne sont pas présentes sur une longue durée, les autres plantes de l'interface vont pouvoir jouer un rôle de refuge viral lorsque l'hôte cultivé disparait (Figure SB.18) (Burdon and Thrall, 2008; Wisler and Norris, 2005).

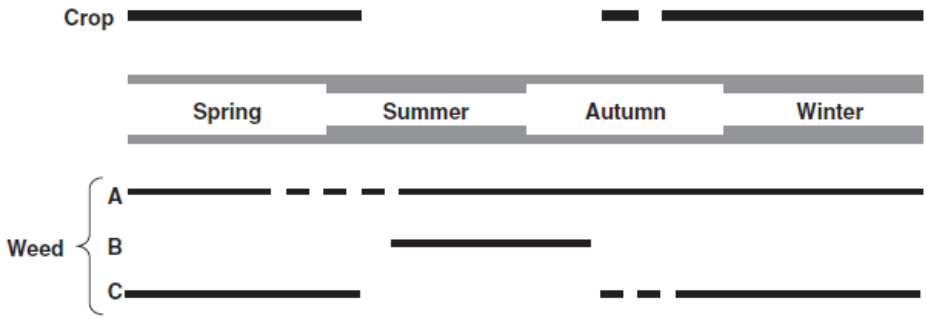


Figure SB.18 : Cycle de vie hypothétique d'un pathogène qui partage des mauvaises herbes et des cultures en tant qu'hôtes. D'après Burdon and Thrall, 2008.

Dans sa revue publiée en 2013, Helen M. Alexander a exposé différentes raisons de s'intéresser aux phytovirus au niveau de cette interface (Alexander *et al.*, 2013) : ils ont souvent de larges gammes d'hôtes ce qui leur permet de couvrir des aires de répartition étendues allant du milieu sauvage au milieu cultivé (Wisler and Norris, 2005) , ils sont dans la majorité des cas transmis par des arthropodes, ce qui augmente leur mobilité dans les paysages (Hogenhout *et al.*, 2008), et ils ont d'importants effets sur les rendements agricoles et sur la fitness des plantes, même en l'absence de symptômes (Cooper and Jones, 2006; Thresh, 2006). Par ailleurs, une étude récente menée dans une interface agro-écologique en Australie a permis de détecter le saut d'hôte du potyvirus *Hardenbergia mosaïc virus* (HarMV) inféodé à son hôte natif *Hardenbergia comptoniana* vers la plante cultivée *Lupinus cosentinii*, ceci amenant à l'hypothèse d'une possible émergence à l'interface entre milieu cultivé et sauvage (Kehoe *et al.*, 2014).

2.3.3.2. Flux de gènes dans l'interface

Contrairement aux plantes présentes dans le milieu sauvage, celles du milieu cultivé présentent des gènes de résistance uniformes (cf. section 2.1.1). L'hybridation introgressive de plantes cultivées avec des plantes sauvages peut amener à l'acquisition de gènes ou d'allèles chez les plantes situées à l'interface (Burdon and Thrall, 2008). Ceci pose un problème pour le contrôle des maladies. En effet, si les gènes de résistance issus du milieu cultivé se retrouvent dans le milieu sauvage, la pression de sélection appliquée sur les pathogènes va avoir tendance à s'uniformiser et des pathotypes plus virulents pourront éventuellement émerger (Burdon and Thrall, 2008). Par analogie, ce problème de transfert horizontal de gènes est bien connu en santé humaine et animale. Les cas de contournement de résistance aux antibiotiques par bactéries de l'environnement ont été extrêmement étudiés.

2.3.3.3. Sélection et diversité dans l'interface

L'interface qui est au carrefour de deux systèmes très différents (Figure SB.19) (Burdon and Thrall, 2013) va être le lieu d'un éventail de pressions de sélection, ce qui doit probablement offrir de nombreuses opportunités évolutives pour les agents pathogènes (Burdon and Thrall, 2008). On peut supposer que l'intensification des pressions de sélection au niveau de l'interface stimule la migration (par exemple celle des insectes vecteurs par des plantes non-hôtes) et la dérive (extinction de population en période défavorable due à des facteurs biotiques). On peut donc formuler l'hypothèse que l'interface est probablement un lieu privilégié pour la diversification génétique des agents pathogènes et de leurs vecteurs, jouant par la même occasion un rôle majeur dans l'émergence des maladies.

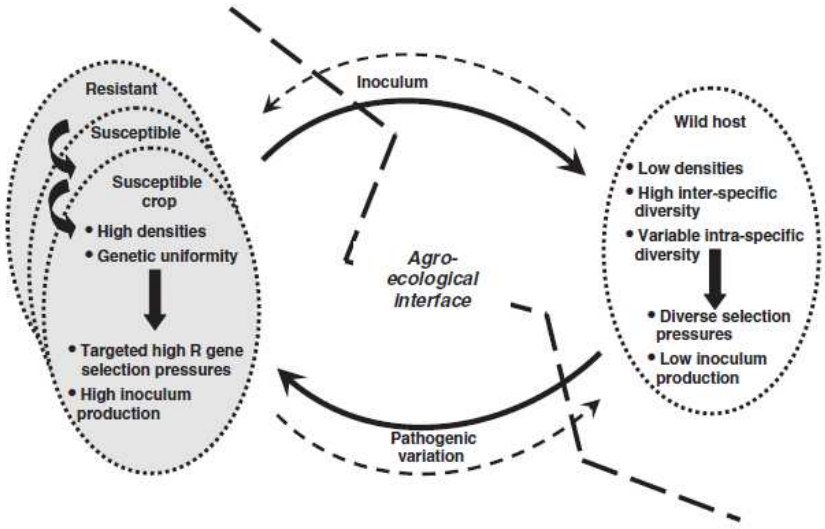


Figure SB.19: Interactions dans l'interface agro-écologique entre les hôtes cultivés et les plantes sauvages (ainsi que les adventices) qui amènent à des échanges d'inoculum et à de la variation des pathogènes. D'après Burdon and Thrall, 2008.

Bien que l'émergence de maladies phytovirales dans l'interface agro-écologique soit encore peu étudiée, les émergences de zoonoses impliquant la destruction des habitats naturels et par conséquent l'augmentation des contacts entre les milieux sauvages et cultivés sont largement documentées. En effet, les maladies telles que la fièvre hémorragique, le *Nipah virus* ou encore la maladie de Lyme sont connues pour avoir émergé au niveau de cette interface créée par l'Homme (reporté dans la revue de Greger, 2007). La consommation de viande provenant d'animaux sauvages par les populations adjacentes aux écosystèmes naturels a également amené à des sauts d'hôtes de leur réservoirs animaux vers l'humain ; un exemple bien connu est celui du HIV (*Human Immunodeficiency Virus*), un virus humain qui aurait évolué à partir du SIV (*Simian Immunodeficiency Virus*) un virus de chimpanzés (reporté dans la revue de Greger, 2007).

2.4. Epidémiologie spatiale ou l'étude des maladies infectieuses dans le paysage

L'épidémiologie phytovirale au sens strict décrit le mouvement d'une maladie virale au sein d'une population de plantes hôtes saines ; cela va donc consister à étudier l'évolution et la distribution spatiale de cette maladie (Alexander *et al.*, 2013; Astier, 2007). Cinq facteurs sont à prendre en compte lorsque l'on s'intéresse à l'épidémiologie des virus de plantes : les propriétés spécifiques du virus (stabilité, gamme d'hôte, symptômes…), l'importance de la source du virus, l'abondance et l'activité des vecteurs, le type de relation entre le virus et les vecteurs, et la sensibilité de la plante hôte (Astier, 2007). Une maladie est souvent agrégée spatialement dans les populations de plantes suite à une transmission vectorielle à courte distance et à l'agrégation des génotypes hôtes similaires (Real, 1996).

2.4.1. Epidémiologie du paysage, un contexte spatio-temporel

2.4.1.1. Définition

La dissémination des parasites est un processus spatio-temporel qui se déroule au niveau de paysages complexes (Burdon and Thrall, 2013). L'épidémiologie du paysage est une branche de l'épidémiologie qui tient compte du fait que les paysages sont des environnements complexes et étudie l'interaction entre l'hétérogénéité du paysage et les processus écologiques sous-jacents qui conduisent à l'apparition, la dissémination et la persistance d'une maladie (Meentemeyer *et al.*, 2012; Plantegenest *et al.*, 2007). Le terme paysage est donc ici utilisé pour décrire un espace hétérogène qui peut influencer les processus de micro-évolution des populations d'agents pathogènes à différentes échelles.

Afin d'évaluer les changements qui se déroulent dans l'espace et dans le temps au sein des populations d'agents pathogènes, des marqueurs moléculaires permettant de caractériser les populations sont très largement utilisés (Nadler, 1995). On parle de « génétique du paysage » lorsque l'on s'intéresse à la distribution géographique des variations génétiques des agents pathogènes et de leurs hôtes et que l'on cherche à décrypter les mécanismes évolutifs et écologiques qui façonnent ces variations génétiques (Biek and Real, 2010). Les agents pathogènes dépendent des ressources hôtes et de leurs vecteurs, et le paysage va donc directement et indirectement affecter la distribution, l'accessibilité et l'abondance de ces ressources. Les variables environnementales vont donc contribuer à l'émergence et à la distribution des agents pathogènes dans l'espace et dans le temps (Ostfeld *et al.*, 2005). La structuration du paysage est donc un des paramètres qui façonnent la structuration génétique des populations de parasites. Par exemple, une structure des hôtes organisée en habitats regroupés (« patch ») peut être déterminante pour la structuration génétique des populations d'agents pathogènes (Biek and Real, 2010). La structuration spatiale va également influencer la coévolution entre les agents pathogènes et leur hôte (Dybdahl and Storfer, 2003). Il est important de préciser qu'un agent pathogène peut également modifier la structuration des hôtes (cf. section 2.2.2) et que les évènements climatiques vont également intervenir dans la distribution des populations d'agents pathogènes (cf. section 2.2.3.5).

2.4.1.2. Aspects spatio-temporels

Le concept de métapopulation (Hanski and Gilpin, 1991) a souligné l'importance de l'organisation du paysage en patchs de populations hôtes sur l'écologie et l'évolution des populations d'agents pathogènes (Burdon and Thrall, 2013). L'ajout du facteur « temps » est important dans ce type d'étude car le facteur « spatial » ne fournit qu'une image figée et les dynamiques virales n'ont pas qu'une dimension spatiale, elles ont aussi une dimension temporelle (Burdon and Thrall, 2013). En effet, les aspects temporels des interactions entre les populations d'agents pathogènes et leurs hôtes jouent un rôle significatif dans l'émergence d'agents pathogènes suite aux modifications

dynamiques de leur pathogénie (Burdon and Thrall, 2013). L'espace et le temps sont donc deux variables fondamentales qui sont sous-jacentes aux forces génétiques qui régissent les directions épidémiologiques et génétiques des trajectoires évolutives (Figure SB.20) (Burdon and Thrall, 2013). Toutefois, peu d'études incorporent à la fois les dimensions temporelles et spatiales dans l'épidémiologie à l'échelle des métapopulations, et la majorité de ces études ne concernent que les champignons phytopathogènes (Burdon and Thrall, 2013).

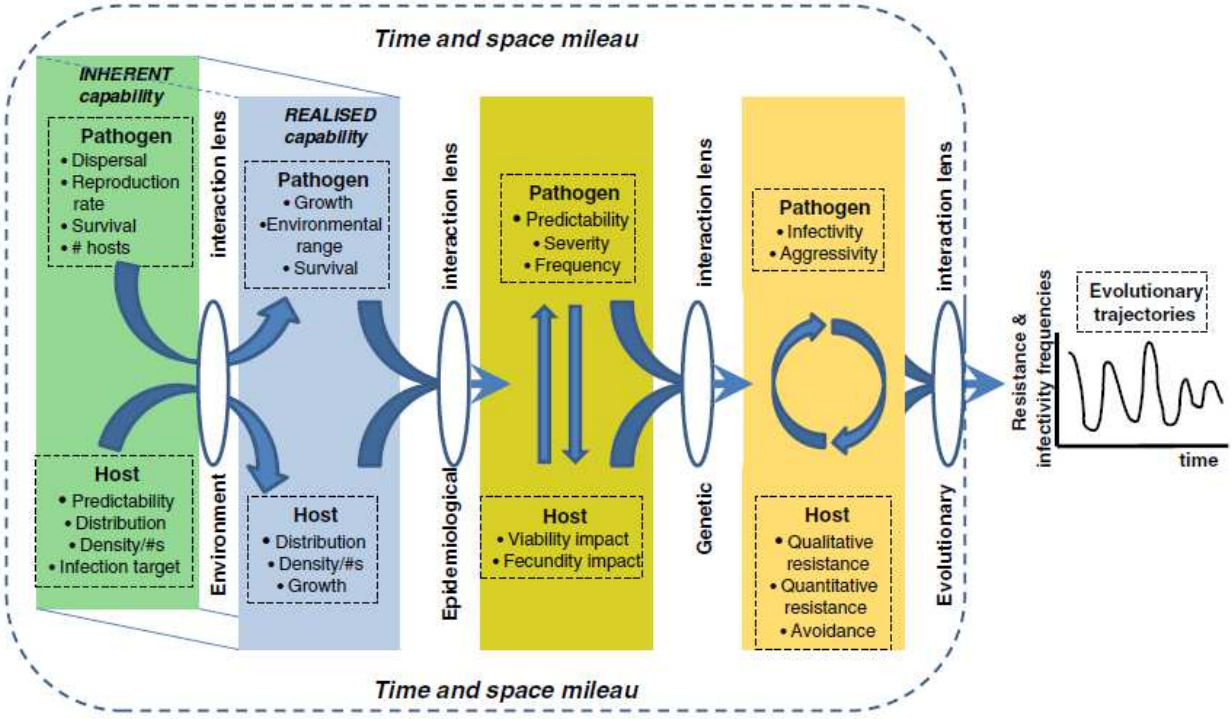


Figure SB.20 : Paramètres écologiques et évolutifs façonnant les trajectoires évolutives des interactions hôtes-pathogènes. D'après Burdon and Thrall, 2013.

2.4.1.3. Connectivité

La connectivité entre les populations influence l'évolution et le maintien de la diversité génétique, des résistances hôtes, et des gènes d'avirulence des parasites (Thrall and Burdon, 2002). Le degré de connectivité entre les hôtes et leurs agents pathogènes est dépendant de leur dispersion, de leur capacité à persister à des stades inactifs, de la structure de l'environnement, et de la compétition interspécifique (Burdon and Thrall, 2013). Toutefois, l'influence de la connectivité sur les paramètres écologiques qui régissent les dynamiques des maladies dans les écosystèmes naturels est difficile à mesurer. Les perturbations de la connectivité se traduisent par l'apparition de dynamiques non-linéaires des variables écologiques, ce qui devrait nécessiter la mise en place d'études à différentes échelles d'observation (Figure SB.21) (cf. section 2.4.14) (Meentemeyer *et al.*, 2012). Ces différences dans la distribution et la connectivité entre les populations ont des impacts sur l'extinction-recolonisation des espèces et leur flux de gènes (Burdon and Thrall, 2013). Tous ces facteurs influencent la probabilité de rencontre entre un hôte et son agent pathogène et par conséquent la co-occurrence des

interactions hôtes/agents pathogènes et les forces de sélections qui leur sont appliquées (Burdon and Thrall, 2013).

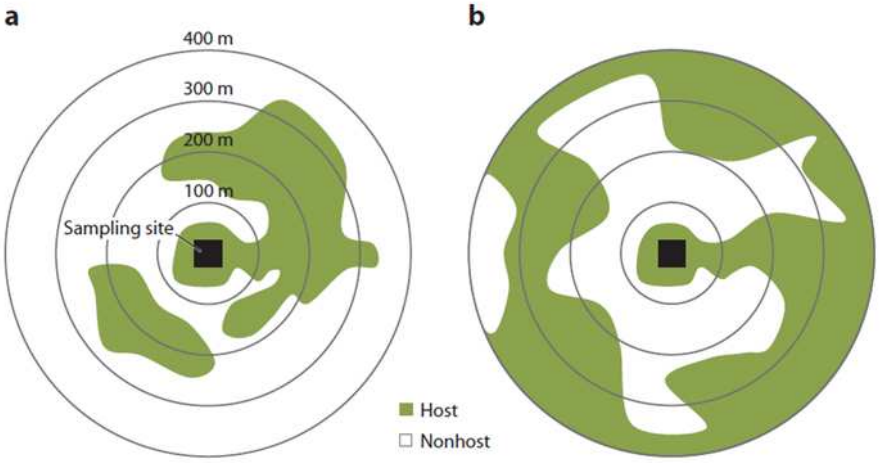


Figure SB.21 : Représentation des effets échelle-dépendants sur l'abondance et la configuration spatiale des espèces hôtes (en vert) et non hôtes (en blanc) au travers d'échelles d'études nichées dont le rayon augmente (incréments de 100m) autour d'un site d'échantillonnage (carré noir). La quantité d'habitat d'espèces hôtes autour de chaque site en conjonction avec la connectivité structurale de l'habitat peut influencer la pression d'infection globale dans un site. Si une analyse avait été conduite à une seule échelle (100m), les scénarios a et b auraient été identiques, alors qu'à 200m les scénarios commencent à se différencier. Si on fait une étude multi-échelle ou une étude à plus grande échelle (e.g. 400m), on se rend compte que le scenario b est représenté par une large aire d'habitats hôtes contigüe et le site d'échantillonnage doit donc subir une plus grande pression d'infection venant du paysage qui l'entoure. D'après Meentemeyer *et al.*, 2012.

2.4.1.4. Echelles d'études

Les agents pathogènes affectent la distribution des plantes à différentes échelles spatiales (Figure SB.22) (Alexander, 2010). Généralement, il est difficile de prédire ce qu'il se passe à une grande échelle via des études à petite échelle. En effet, les composantes hétérogènes d'un habitat et de l'environnement associés à différentes échelles spatiales peuvent affecter les patrons des épidémies (Archie *et al.*, 2009). Il y a rarement une seule échelle correcte pour étudier un pathosysthème (Meentemeyer *et al.*, 2012) et l'échelle à laquelle sont faites les observations peuvent influencer et biaiser les conclusions (Chave, 2013; Plantegenest *et al.*, 2007). En effet, des processus épidémiologiques peuvent opérer de manière différente selon l'échelle à laquelle on les observe (Meentemeyer *et al.*, 2012). Par exemple, une étude menée en 2012 s'est intéressée à la prévalence du B/CYDV suivant deux échelles, l'une (la prairie) nichée dans l'autre (complexe de prairies). Cette étude a démontré que la prévalence du B/CYDV varie en fonction de la prairie, *a contrario* la prévalence ne varie pas entre les différents complexes de prairies (Moore and Borer, 2012). Habituellement, la majorité des études se focalisent sur une seule échelle : sur 143 études d'épidémiologie du paysage, seulement 13% utilisent une approche multi-échelle (Meentemeyer *et al.*, 2012).

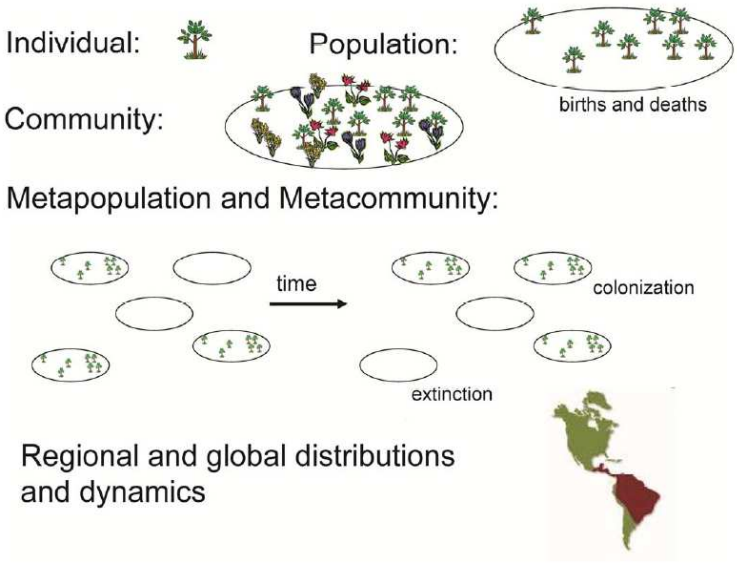


Figure SB.22 : Différentes échelles d'étude en épidémiologie du paysage. Les études de métapopulations considèrent plusieurs populations dans le paysage. Des évènements d'extinction et de colonisation locaux peuvent advenir au niveau des populations individuelles. Les métacommunautés font référence au même concept mais au niveau de la communauté (c'est-à-dire qu'on va considérer plusieurs espèces). D'après Alexander, 2010.

Le choix de l'échelle dépend de l'étude que l'on veut mener. Les grandes échelles (régionale à continentale) peuvent être utilisées pour étudier l'impact potentiel d'agents pathogènes émergents à large échelle alors que les études à plus petites échelles peuvent être menées pour connaître l'hétérogénéité spatiale de la disponibilité en hôte et comment elle influence l'expansion de la maladie (Meentemeyer *et al.*, 2012). Bien évidemment, le choix de l'échelle est contraint par le temps et les ressources dont on dispose. Ross Meentemeyer préconise l'analyse multi-échelles qui consiste à collecter des données et utiliser des méthodes analytiques sur au moins deux niveaux d'intérêts (Meentemeyer *et al.*, 2012).

2.4.2. Mener une étude d'épidémiologie du paysage

2.4.2.1. Échantillonnage

L'échantillonnage est la première étape dans les études d'épidémiologie spatiale. Les pathogènes étant généralement répartis de façon hétérogène dans le paysage il est important de mettre en place un dispositif adéquat au type d'étude envisagé concernant la taille de l'échantillon et le type d'échantillonnage.

2.4.2.2. Taille de l'échantillon

La détermination de l'effectif de l'échantillon à étudier est cruciale. Dans les études de diversité, il est important d'avoir un échantillon représentatif de la communauté de manière à répondre à la question scientifique posée avec une précision suffisante en fournissant un effort d'échantillonnage minimal. Ainsi, dans une étude de diversité, pour déterminer si un échantillonnage est représentatif de la communauté étudiée, on utilise souvent les courbes de raréfaction. Cette courbe est construite sur la

base d'un sous-échantillonnage de l'inventaire complet des effectifs et en calculant le nombre d'espèces présentes dans chacun de ces échantillons. Ainsi en abscisse est représentée la taille des sous échantillons et en ordonnée le nombre d'espèces. Quand cette courbe atteint un plateau on estime que l'échantillon est représentatif de la communauté en terme de diversité.

2.4.2.3. Stratégies d'échantillonnage

Il existe 3 grandes catégories d'échantillonnage : aléatoire, systématique et stratifié. Avant de décrire ces stratégies il est important de définir ce que sont un quadrat et un transect. Etablir des quadrats consiste à déterminer des aires délimitées de taille plus petites que la zone globale d'échantillonnage dans lesquels on va effectuer l'échantillonnage ; les quadrats sont donc des unités d'échantillonnage. Lorsqu'on établit une ligne virtuelle dans l'aire d'échantillonnage, on parle de transect ; les plantes sont échantillonnées sur des points ou des aires régulières le long de cette ligne. Cette méthodologie peut être implémentée dans l'établissement de quadrats.

- ***Echantillonnage aléatoire***

L'échantillonnage aléatoire est la méthode la moins biaisée des 3, et n'implique aucune subjectivité, chaque membre de la population a la même chance d'être sélectionné. Trois types de dispositifs peuvent être mis en place afin d'effectuer un échantillonnage aléatoire (Figure SB.23) :

- Echantillonnage sur des points aléatoires : une grille peut être placée sur la carte représentant l'aire d'étude, des tables de nombres aléatoires vont être utilisées pour obtenir les coordonnées de chaque point d'échantillonnage sur la grille. Les échantillons récoltés doivent alors être les plus proches des points ainsi définis.

- Echantillonnage aléatoire sur une ligne : une paire de points aléatoires est positionnée sur une grille, ces deux points vont être reliés par une ligne sur laquelle on va mener l'échantillonnage.

- Aire d'échantillonnage aléatoire : à partir du même type de grille, on va générer les coordonnées du point inférieur gauche d'une aire d'échantillonnage via une table de nombres aléatoires.

- ***Echantillonnage systématique***

Les échantillons sont choisis de manière systématique, en effet ils vont être régulièrement distribués dans l'espace (par exemple tous les 500m sur une ligne d'échantillonnage). Ici aussi trois types de dispositifs peuvent être mis en place (Figure SB.23):

- Echantillonnage systématique sur points : les points d'échantillonnage sont définis sur une grille placée sur la carte d'échantillonnage. Les points sont placés à l'intersection des lignes ou au centre de chaque carré de la grille.

- Echantillonnage systématique en ligne : on peut positionner une ligne de transect sur une carte allant par exemple de l'Ouest à l'Est, et échantillonner sur des points équidistants le long de cette ligne (par exemple tous les 500m).

- Aire d'échantillonnage systématique : l'échantillonnage est effectué dans des carrés d'une grille placée sur la carte. Les carrés sont choisis selon une régularité qui peut être par exemple un carré sur trois de gauche à droite et de haut en bas (Fig. SB.23).

- ***Echantillonnage stratifié***

Lorsque la population que l'on veut échantillonner peut être divisée en sous-catégories représentatives (Figure SB.23) on parle d'échantillonnage stratifié. Ce type d'échantillonnage compte deux types de dispositifs :

- Echantillonnage stratifié systématique : la population est divisée en groupes connus et chacun des groupes est échantillonné de manière systématique (décrit précédemment).

- Echantillonnage stratifié aléatoire : ici chacune des sous-catégories est échantillonnée de manière aléatoire (décrite précédemment).

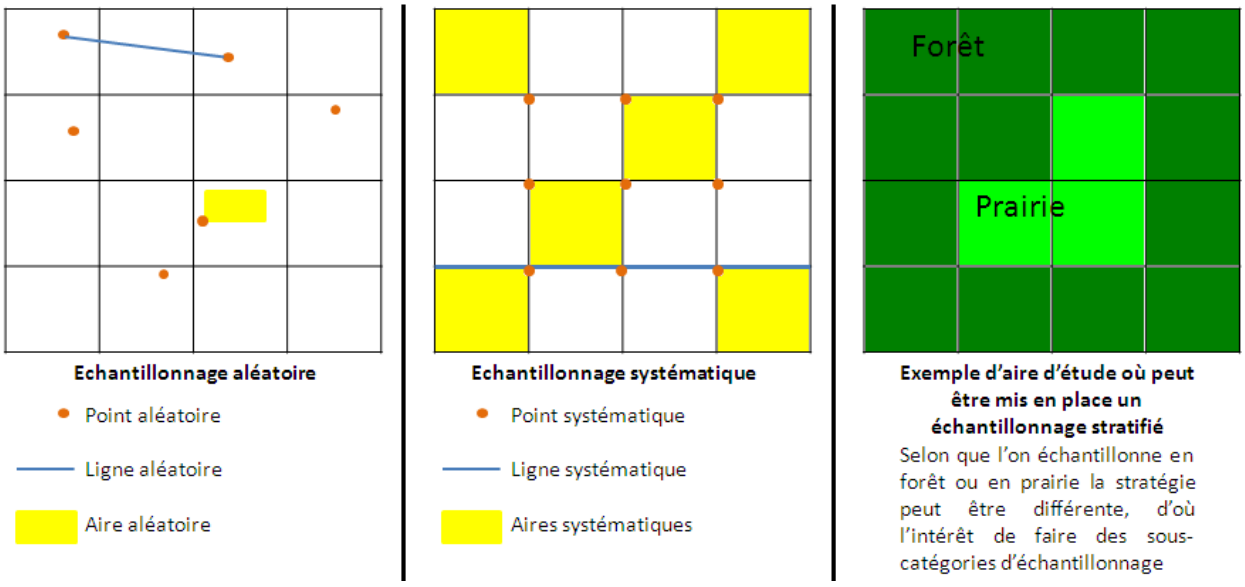


Figure SB.23 : Illustration des différentes stratégies d'échantillonnage.

2.4.2.4. Récolte de variables

En épidémiologie du paysage on cherche à comprendre l'occurrence et la progression d'une maladie. Pour cela, il faut chercher les variables qui vont pouvoir expliquer les patrons spatiaux des parasites. Ces patrons spatiaux des parasites sont souvent expliqués par les patrons spatiaux de leurs hôtes (eux-mêmes expliqués par certaines variables). Outre les variables intrinsèques aux plantes hôtes (traits d'histoire de vie), on va pouvoir corréler les patrons à des variables environnementales. Bien que certaines des variables n'aient besoin d'aucun outil particulier pour être déterminés (par exemple le fait qu'une plante soit cultivée ou non), différents outils de collecte de

variables sont aujourd'hui au point permettant d'avoir des informations supplémentaires sur l'aire d'échantillonnage.

Une des grandes catégories de ces outils est celle de la télédétection qui contient l'imagerie spectrale/thermique/hyperspectrale/multispectrale qui permet d'obtenir des informations qu'on ne peut voir à l'œil nu. À titre d'exemple, l'imagerie hyperspectrale va permettre de détecter des symptômes qu'on ne peut voir à l'œil nu dans des céréales infectées par du BYDV (Jones, 2013), ou encore un dispositif de capture d'image a permis de différencier les degrés d'infection du MSV dans des feuilles de maïs (Martin *et al.*, 1999). A échelle plus large, l'utilisation combinée de la télédétection aérienne et de prise de données au sol permettent de maximiser la précision et l'exactitude de l'évaluation des maladies et leur quantification (Steddom *et al.*, 2005). Des échelles régionale à continentale, des images générées par satellites vont dans certains cas permettre d'identifier des cultures infectées par des virus ; c'est surtout le cas de grandes cultures pour lesquels les symptômes sont manifestes (par exemple des cultures de blé infectées par le *Wheat streak mosaic virus* on pu être détectées via des images satellites au Texas (Mirik *et al.*, 2011)).

En plus de récolter des variables sur les épidémies on peut également récolter des variables météorologiques (température, humidité, taux de CO_2), édaphiques (type de sol, pH…), sur le type de végétation et toute autre variable qui pourrait avoir une influence sur les patrons des maladies observés dans le paysage. Ces données récoltées via les systèmes d'information géographique (SIG) peuvent alors être intégrées dans des analyses spatiales statistiques puis être à nouveau projetées sur une carte (Kitron, 1998). On citera également le « Global Positioning System » (GPS) qui va permettre de positionner les variables et les points d'échantillonnage de manière précise sur une carte.

2.4.2.5. Statistiques spatiales

Les statistiques spatiales vont permettre de décrire, expliquer, extrapoler, et prédire la distribution des objets et des processus dans l'espace (Kitron, 1998). Des méthodes non-géographiques ont longtemps été utilisées par les écologistes, elles se focalisaient sur des mesures d'agrégation basées sur des comptages de fréquence des individus dans les quadrats sans tenir compte de leur positionnement géographique (Pielou, 1969). Certaines statistiques spatiales géographiques telles que la « nearest neighbor method » ne tiennent compte que de la localité alors que les autres vont considérer à la fois la localité et les valeurs des variables qui lui sont associées (Bailey, 1994). Ces dernières sont les suivantes (Kitron, 1998):

- Les mesures d'autocorrélation spatiale peuvent être utilisées pour étudier la distribution spatiale statique des épidémies et de leur vecteur. Ces mesures révèlent la tendance de localités proches à s'influencer les unes les autres plus que les localités éloignées. L'analyse de l'autocorrélation permet de quantifier la régularité spatiale d'un phénomène (une forme de complexité spatiale) et de déterminer la portée de la

dépendance spatiale afin, notamment, de définir un dispositif d'échantillonnage garantissant l'indépendance des données (Meentemeyer *et al.*, 2012).

-Les géostatistiques vont permettre l'étude de variables régionalisées. Parmi les méthodes utilisées en géostatistiques, les semivariogrammes vont permettre l'analyse de variables spatialement distribuées et l'estimation de valeurs pour les localisations non échantillonnées (Nelson *et al.*, 1999). À partir du semivariogramme, le krigeage (kriging) aussi appelé interpolation spatiale va permettre d'interpoler des valeurs sur des sites non échantillonnés via le calcul de l'espérance mathématique. Le krigeage tient compte à la fois de la distance entre les données et le point d'estimation et des distances entre les données deux à deux.

-Les techniques analytiques permettant de combiner les statistiques spatiales avec des analyses temporelles telles que l'autogression spatiale. Elles vont permettre une analyse complète de l'épidémiologie du paysage. Il existe une grande diversité de méthodes permettant ce type d'analyse (http://www.stat.unc.edu/faculty/rs/s321/spatemp.pdf) (Mardia and Goodall, 1993). À titre d'exemple, la méthode SADIE permet l'analyse de la conservation de patrons spatiaux au cours du temps via des indices de distances. Les patrons spatiaux vont être matérialisés sur des cartes et comparés à différents temps (Perry, 1999).

Les méthodes de statistiques spatiales peuvent être utilisées en combinaison avec des statistiques classiques telles que des régressions, des analyses de variances, voire des analyses multivariées (Kitron, 1998; Turner, 2005).

3. La métagénomique virale

3.1. Métagénomique virale et pathologie : Une histoire récente

Les approches de métagénomique nées à la fin des années 1980 apparaissent comme un outil de choix pour l'étude des communautés virales. Cette partie a fait l'objet d'une revue d'articles qui a été publiée en mai 2013 : Bernardo, P., Albina, M., Eloit, M., Roumagnac, P. (2013). Métagénomique virale et pathologie, une histoire récente. Médecine et sciences : M/S 29(5), 501-508.

Cette publication retrace l'émergence de la métagénomique en pathologie et fait un état de l'art de son utilisation en virologie pour la découverte de nouveaux virus, leur diagnostic, leur évolution intra-hôte ainsi que l'épidémiologie moléculaire des maladies virales.

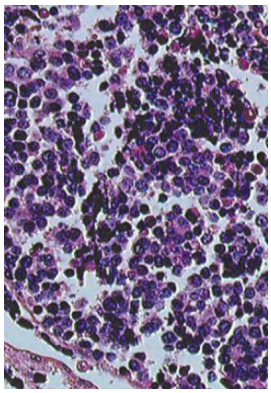
médecine/sciences 2013 ; 29 : 501-8

Métagénomique virale et pathologie

Une histoire récente

Pauline Bernardo[1], Emmanuel Albina[2,3], Marc Eloit[4], Philippe Roumagnac[1]

> L'étude des maladies virales humaines, animales et végétales a récemment bénéficié du développement de la métagénomique en écologie. Cette approche, dite sans *a priori*, consiste à décrire le génome viral total, ou virome, d'une fraction d'un écosystème sans tenir compte de cibles moléculaires déjà connues. Les premiers résultats de la métagénomique en pathologie virale ont été obtenus très localement au niveau d'un tissu ou d'un organe. Ils ont débouché sur la découverte de nouveaux virus et ont permis des avancées sur le plan du diagnostic, de l'épidémiologie moléculaire et de l'évolution virale. Ces travaux sont d'une grande pertinence pour réévaluer des concepts en pathologie et, notamment, le rôle biologique des virus dans un organisme. Ils révèlent la nécessité d'actualiser les méthodes de diagnostic. <



Vignette (Photo © Jean Kanitakis).

[1] Cirad (Centre de coopération internationale en recherche agronomique pour le développement), UMR BGPI (biologie et génétique des interactions plante-parasite), 34398 Montpellier Cedex 5, France ;
[2] Cirad, UMR CMAEE (contrôle des maladies animales exotiques et émergentes), 97170 Petit-Bourg, Guadeloupe, France ;
[3] Inra, UMR1309 CMAEE, 34398 Montpellier, France ;
[4] Institut Pasteur, laboratoire de découverte de pathogènes, département de virologie, 28, rue du Docteur Roux, 75015 Paris, France.
philippe.roumagnac@cirad.fr

La métagénomique microbienne et virale : une nouvelle approche en écologie microbienne

La métagénomique (voir *Glossaire*) est un terme qui est apparu en 1998 dans un article écrit par Jo Handelsman *et al.* [1]. Il faut toutefois remonter au début des années 1990 pour trouver trace des premiers travaux de métagénomique [2], et même aux années 1980 pour deviner l'apparition de ce nouveau concept grâce aux travaux de N.R. Pace *et al.* [3] qui proposent pour la première fois de cloner directement les gènes 5S et 16S des ARN ribosomaux (ARNr) à partir d'échantillons prélevés dans l'environnement. Le terme métagénomique est ensuite peu à peu devenu le terme consacré pour désigner les travaux de génomique environnementale, génomique des communautés, écogénomique, génomique des populations microbiennes, etc. Les études engagées à la fin du XX[e] siècle avaient toutes pour objectif de dépasser le cadre des collections de souches de microorganismes des laboratoires en séquençant directement les acides nucléiques collectés *in situ* dans différents écosystèmes [4]. Ce nouvel angle de recherche permettait de s'affranchir de l'isolement et de la culture des souches bactériennes et, *de facto*, d'avoir accès aux acides nucléiques des souches non cultivables, souches qui représentent plus de 99 % de la biodiversité des micro-organismes [5, 40]. Cette nouvelle approche en écologie microbienne visait ainsi à mieux décrypter la diversité des microorganismes présents au sein d'une niche écologique, à mieux comprendre les relations évolutives des différents taxons identifiés et à caractériser de nouveaux gènes et de nouvelles fonctions [4]. Craig Venter et son équipe [6] ont démontré de façon spectaculaire la puissance de cette nouvelle approche en identifiant 148 phylotypes et 1,2 millions de nouveaux gènes par la seule analyse de quelques échantillons prélevés dans la mer des Sargasses.

À la suite des travaux pionniers de T. Allander *et al.* [7], la métagénomique virale émerge en 2002 avec un papier fondateur de M. Breitbart, *et al.* [8]. L'approche utilisée, dite sans *a priori*, consiste à purifier l'ensemble des particules virales d'un échantillon d'un écosystème ciblé puis, indépendamment de cibles moléculaires déjà connues (approche indépendante de séquence), à amplifier le génome viral total (virome) en utilisant des amorces aléatoires. L'ensemble de ces études, décrites dans environ une trentaine d'articles publiés fin 2012, met en évidence un nombre considérable de séquences d'ADN inconnues (> 50 %) ne correspondant à aucune donnée actuellement archivée dans les bases de données internationales [9]. Cette « matière noire » représente le plus important réservoir d'information génétique encore non décodée au sein de la biosphère. De plus, quand les séquences en acides aminés peuvent être décodées, elles présentent le plus souvent des similitudes

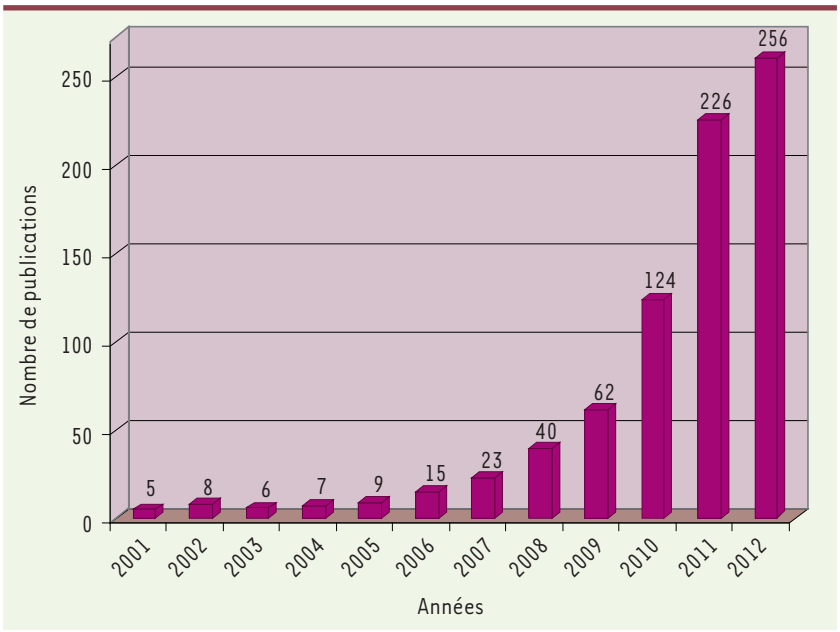


Figure 1. ***Nombre de publications en métagénomique enregistrées annuellement par le NCBI (national center for biotechnology) depuis 2001.*** Les recherches ont été faites avec les mots-clés suivants : *virus AND metagenomics OR next generation sequencing OR high throughput sequencing*.

très faibles (< 50 %) avec des séquences déjà connues et caractérisées [9]. Les virus environnementaux sont plus ou moins proches des virus actuellement connus. Ils sont généralement rattachés à de nouvelles espèces virales qui, à leur tour, permettent de densifier les arbres phylogénétiques des familles ou des genres viraux [10].

L'ensemble de ces premiers résultats de métagénomique virale et microbienne a contribué à remettre en cause plusieurs concepts communément admis en microbiologie.

Le premier concept est d'ordre ontologique. Ces études révèlent la complexité de la diversité globale des génomes présents au sein des écosystèmes naturels, avec nombre de génomes chimériques et des populations à forte hétérogénéité génétique. Les concepts de génomique horizontale [11] et de métaorganismes [12] sont en train d'émerger ; ils se fondent sur l'idée suivante : un métagénome peut être considéré comme une ressource génomique commune, partagée par un ensemble d'entités aux contours plastiques (le métaorganisme) [12]. Le microbiome humain est sûrement le métaorganisme le plus connu : les cellules bactériennes qui le composent sont dix fois plus nombreuses que les cellules du corps humain les hébergeant [13]. Il a été récemment montré que le microbiome du tube digestif humain est le siège d'échanges génétiques, ainsi que de phénomènes de régulations, de mutualisme et de compétition [14]. La question de la définition d'un être pluricellulaire, l'hôte *stricto sensu* ou bien l'hôte et son cortège de microorganismes, a donc été posée pour essayer de mieux comprendre l'adaptation de l'hôte à son environnement [12].

Le deuxième concept que l'on peut interroger est d'ordre à la fois écologique et pathologique. Les microbes au sens large, et les virus en particulier, sont-ils nécessairement dangereux, en particulier pour la santé de leurs hôtes eucaryotes [15] ? Plusieurs études récentes ont montré que des micro-organismes peuvent être, soit parasitaires, soit mutualistes, en fonction des paramètres environnementaux. À titre d'exemple, il a été publié que des staphylocoques de la flore commensale de l'épithélium cutanée jouent un rôle clé dans la modulation de l'inflammation déclenchée par une blessure cutanée [16].

Les interrogations relatives à ces deux concepts, et surtout le nombre croissant de nouveaux microorganismes découverts par des approches sans *a priori*, ont conduit à l'adoption de la métagénomique par une partie des pathologistes au début du XXIe siècle. L'idée générale a été d'appliquer les approches de métagénomique à la pathologie en déplaçant le cadre d'étude de l'écosystème à l'organisme. Le terme de métagénomique est donc ici légèrement déformé par rapport à son acception première, et certains scientifiques préfèrent utiliser le terme de « nouvelles techniques de séquençage » (NextGEN ou NGS, *next generation sequencing*) pour nommer cette nouvelle approche. Nous garderons ici le terme de métagénomique appliquée à la pathologie car l'utilisation des nouvelles techniques de séquençage se trouve ici à nouveau couplée à des méthodes d'extraction et d'amplification sans *a priori*. Ces travaux ont particulièrement fleuri en virologie, ce qu'illustre la croissance exponentielle des publications depuis 2005 *(Figure 1)*. Nous nous concentrerons donc dans cette revue sur les principaux résultats obtenus depuis 2001 et les travaux pionniers de T. Allander *et al*. [7] en métagénomique virale.

Les premières études basées sur la métagénomique virale ont mis l'accent sur la découverte de nouveaux virus [10]. Mais depuis, plusieurs avancées ont été réalisées dans d'autres disciplines de la virologie grâce aux résultats issus de travaux de métagénomique virale et, plus largement, de l'utilisation des NGS. Nous discuterons ici de ces avancées dans quatre champs : la découverte de nouveaux virus, le diagnostic, l'épidémiologie moléculaire et l'évolution *(Figure 2)*.

Métagénomique virale et découverte de nouveaux virus

Dès 2008, un papier *princeps* a illustré le potentiel des approches métagénomiques de criblage à haut débit (*high-throughput screening*, HTS) en identifiant le

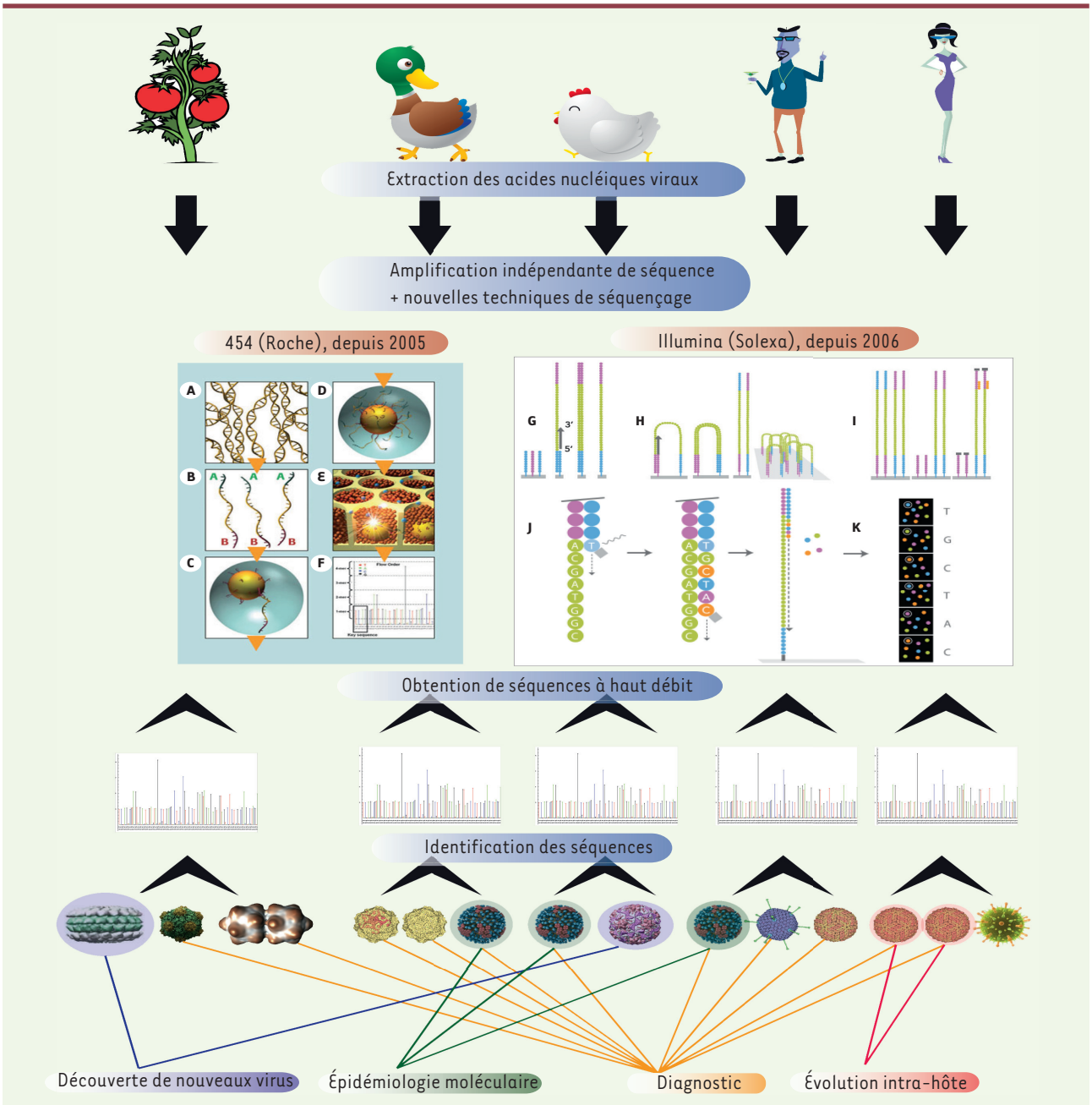


Figure 2. ***Principales étapes des travaux de métagénomique.*** Les techniques de séquençage à haut débit (454 [Roche] et Illumina [Solexa]) sont de nos jours majoritairement utilisées dans ces travaux. Pour la technique 454 (Roche), six étapes ont lieu successivement. ***A, B.*** Construction d'une librairie avec dénaturation de l'ADN et fixation d'adaptateurs ; ***C.*** Fixation de l'ADN sur des microbilles contenant à leur surface des oligonucléotides complémentaires aux adaptateurs ; ***D.*** Emulsion PCR ; ***E.*** Pyroséquençage dans des plaques en fibre optique picotitrée. Chaque nucléotide incorporé dans la synthèse du brin complémentaire libère un groupe pyrophosphate qui engendre la production d'un signal lumineux. ***F.*** Lecture des signaux lumineux aboutissant à la production d'un pyrogramme indiquant la séquence obtenue. La technique Illumina (Solexa) se déroule en cinq étapes successives. ***G.*** Dénaturation de l'ADN, fixation d'adaptateurs, puis hybridation des fragments d'ADN avec les oligonucléotides complémentaires des adaptateurs sur un support ; ***H.*** Les brins se courbent et s'hybrident avec les oligonucléotides proches. Le brin complémentaire est synthétisé par PCR. ***I.*** Les ADN double-brins sont dénaturés et les processus ***G*** et ***H*** reprennent ; ***J-K.*** L'incorporation de nucléotides marqués par des fluorochromes durant la synthèse d'ADN permet une lecture de la séquence. Les séquences obtenues à haut débit sont identifiées par similarité de leurs séquences (Blast, *basic local alignment search tool*) avec celles d'une base de données internationale (GenBank, NCBI). Les résultats de ces identifications de séquences sont exploitables au sein de quatre disciplines de la virologie : (1) la description de nouveaux virus qui implique des travaux de systématique et de taxonomie, (2) l'épidémiologie moléculaire avec pour corollaire la surveillance virale, (3) le diagnostic viral et (4) l'évolution virale.

GLOSSAIRE

Contig : séquence obtenue à la suite de l'assemblage de séquences contiguës présentant des zones identiques et chevauchantes à leurs extrémités 3' et 5'.

Écogénomique : méthode qui consiste à « étiqueter » les séquences de chaque échantillon avant le séquençage de manière à pouvoir réattribuer chaque séquence à son échantillon d'origine à la suite du séquençage à haut débit.

Métagénomique : analyse sans étape de culture de l'ensemble des acides nucléiques des microorganismes présents dans un milieu donné. Le catalogue de ces séquences représente le génome collectif ou le génome global d'un échantillon donné, appelé métagénome.

Microbiome : ensemble des génomes des bactéries colonisant l'organisme d'un animal.

NGS-HTS (*next generation sequencing/high throughput sequencing*) : nouvelles techniques de séquençage permettant la production d'un grand nombre de séquences rapidement et à moindre coût par comparaison avec la méthode Sanger (méthodes post-Sanger). Elles permettent notamment le séquençage d'un mélange de séquences présentes dans un échantillon.

***Pipeline* :** procédé qui permet de lier et automatiser de façon indépendante les différentes étapes de traitement d'une instruction par le processeur comme par exemple l'ensemble des traitements bioinformatiques successifs visant à analyser les jeux de données issus du séquençage à haut débit.

***Reads* :** lectures de fragments nucléotidiques. Par exemple, un cycle de 454 FLX délivre environ un million de lectures de 700 pb chacune en 20 h, soit 900 Mb de données.

Virome : ensemble des génomes d'une population virale trouvés dans un même organisme ou dans un même environnement.

premier polyomavirus oncogène humain, le virus de Merkel [17]. Dans la tumeur de Merkel (une tumeur d'une cellule du système neuroendocrine à la base de l'épiderme), des transcrits codant pour une protéine similaire à l'oncogène T du virus SV40 ont été identifiés [41]. Le virus découvert sur cette base est présent à la surface de la peau chez 70 % des individus, mais, chaque année, seuls un à trois individus sur un million développeront un cancer. Un des mérites de ce travail est de démontrer que des virus extrêmement fréquents peuvent rester inconnus longtemps et être révélés par des approches de ce type. À ce jour, c'est le seul nouveau virus oncogène détecté par ce type d'approche et, *a posteriori*, on ne peut que s'étonner qu'il ait été aussi le premier virus humain découvert par ce type de technologie.

La découverte de virus inconnus est rendue difficile par le fait qu'il n'existe pas de séquences universellement conservées chez les virus. Cet écueil méthodologique a probablement eu pour effet de stimuler les approches indépendantes de séquences en pathologie virale. Pour découvrir des virus inconnus, il faut par ailleurs être capable d'assembler *de novo* de longs « contigs » (blocs de séquences continues, voir *Glossaire*) au sein d'échantillons biologiques complexes. En effet, seuls des contigs de taille relativement grande peuvent permettre d'identifier comme significatifs des pourcentages d'identité très faibles avec des virus connus, voire d'identifier des motifs génomiques indépendants des homologies de séquence. On peut néanmoins prédire que le nombre de virus distants découverts va augmenter avec le développement récent de séquenceurs conjuguant haut débit et longs *reads* (lectures de fragments de séquence nucléotidique, voir *Glossaire*), de manière à pouvoir identifier des virus distants malgré leurs titres faibles.

Pour l'instant, c'est dans son application aux infections virales productives que ce type d'approche s'est révélé le plus performant. Par exemple, l'étude du virome fécal de nombreuses espèces a permis d'identifier un nombre croissant de virus nus, particulièrement au sein de la famille des *Circoviridae*. La résistance de ces virus et leur excrétion avec un fort titre dans un matériel biologique qu'il est possible de concentrer relativement facilement sont pour beaucoup dans cette accumulation de connaissances [18]. Nous avons ainsi identifié une nouvelle espèce virale dans la famille des *Picornaviridae*, qui définit un nouveau genre viral que nous avons appelé *Pasivirus* (*Parecho Sister Clade*), compte tenu de sa proximité avec les parechovirus [19]. De même, à la surface de la peau humaine, notre équipe a identifié différents virus au sein des *Polyomaviridae* (HPyV9) [20] et *Circoviridae* (genre *Gyrovirus*) [21]. La flore virale des poumons [22] et celle de la peau [23] se sont également avérées très riches *(Figure 3)*. Dans la plupart de ces études qui concernaient des syndromes où ces virus ont été recherchés *a posteriori*, aucune relation ultérieure de ce virus avec une maladie n'a pu être mise en évidence.

Au-delà du cas des infections inapparentes, et hormis le cas remarquable du virus de Merkel évoqué plus haut, l'identification de virus responsables de maladies s'est révélée beaucoup plus facile dans les maladies aiguës. C'est le cas de l'identification du virus de Schmallenberg, un nouvel orthobunyavirus émergeant en Europe chez les ruminants [24], du picornavirus de l'hépatite de la dinde [25] et de l'arterivirus hémorragique simien [26]. À l'inverse et jusqu'à présent, le virus de Merkel reste le seul virus identifié par NGS responsable de maladies chroniques incluant les cancers. De manière générale, les implications biologiques et thérapeutiques de la découverte fortuite ou orientée de séquences virales dans un syndrome infectieux nécessiteront un travail d'imputabilité, qui doit reposer sur une actualisation des postulats de Koch[1] que certains ont pu proposer [27]. Cette imputabilité pourrait être difficile à établir pour des infections persistantes n'évoluant vers des conséquences pathologiques que chez un petit nombre d'individus.

[1] Postulats de Koch : (1) l'agent doit être présent chez tous les individus atteints ; (2) l'agent peut être cultivé ; (3) l'agent introduit chez l'hôte, ou dans un modèle animal proche, induit la maladie ; (4) l'agent peut être à nouveau isolé à partir d'un animal présentant la pathologie.

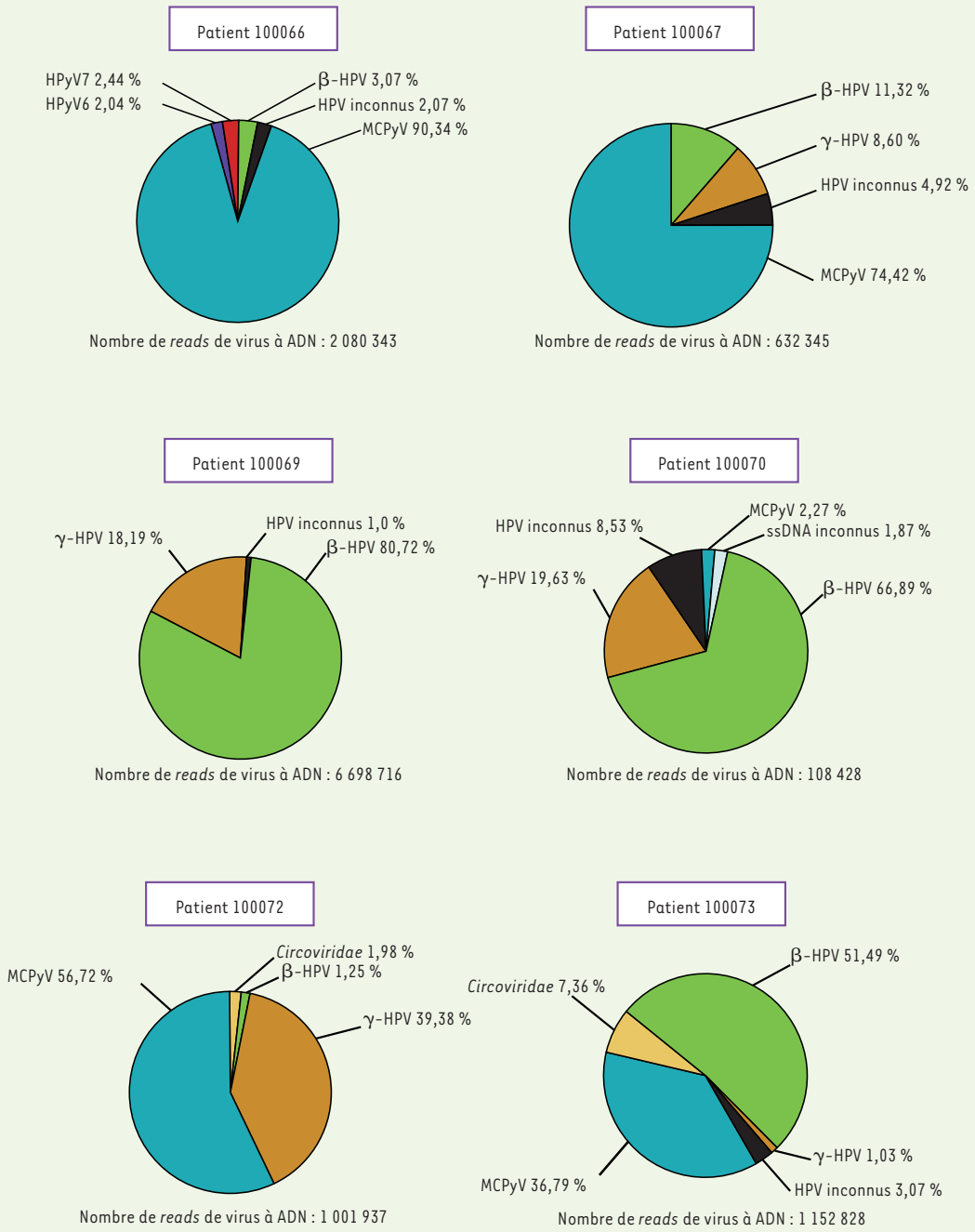


Figure 3. ***Analyse métagénomique de la flore virale au niveau de la peau.*** L'abondance relative des lectures (*reads*) de papillomavirus (β-HPV, γ-HPV et HPV inconnus), de circovirus et d'autres virus à ADN simple brin, ainsi que des polyomavirus (MCPyV, HPyV6 et 7), a été mesurée sur la peau de six patients dont cinq patients sains et un patient malade présentant une tumeur de Merkel due au virus de Merkel (MCPyV) (patient 100066) [23]. ssDNA : *single-stranded DNA*.

Application de la métagénomique virale au diagnostic

L'utilisation des NGS dans une approche diagnostique apparaissait jusqu'à il y a peu comme totalement irréaliste en raison de la combinaison de coûts élevés et d'une durée longue de séquençage et d'analyse. Cette perception est en train de changer pour plusieurs raisons. L'apparition de séquenceurs de paillasse (Illumina MiSeq, Ion Torrent PGM) réduit la durée de séquençage à moins de 24 h au prix néanmoins d'une diminution du débit sans doute critique pour un certain nombre d'indications. Encore plus récemment, le Proton de Ion Torrent a été mis sur le marché : il est doté d'un

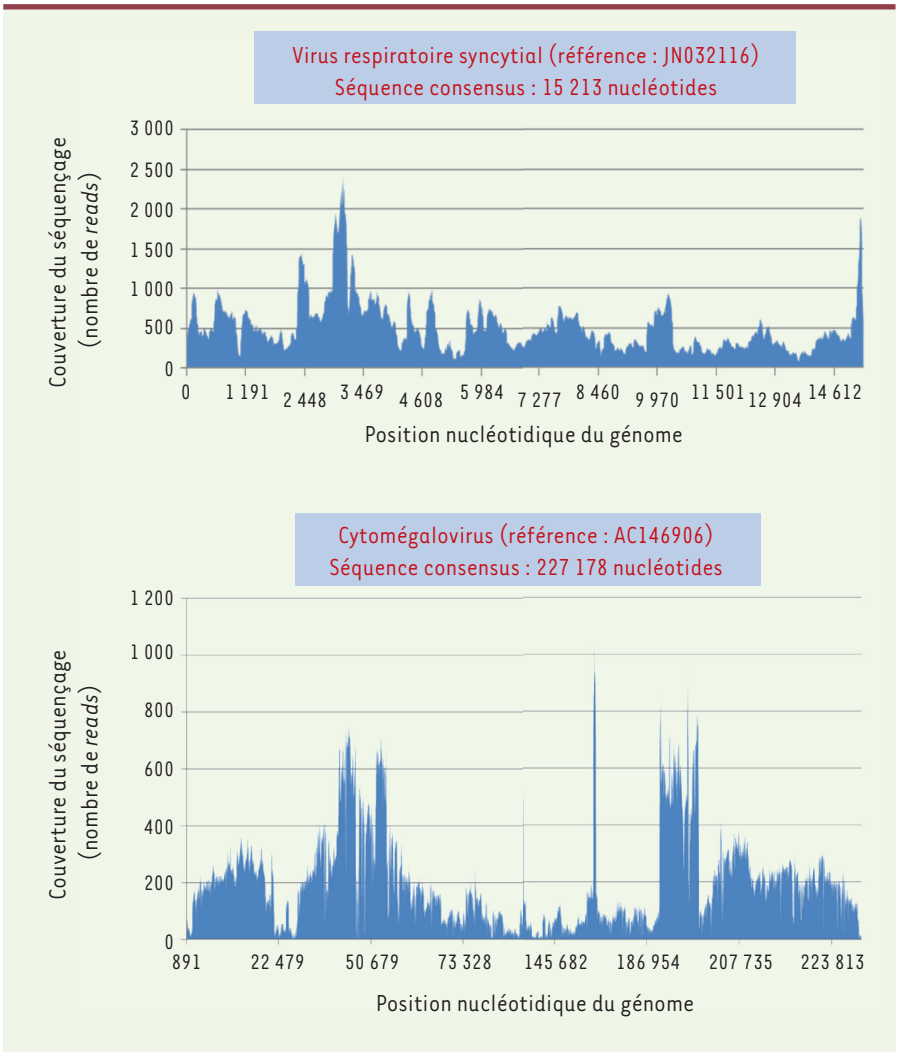


Figure 4. ***Analyse d'un écouvillonnage nasal par la technologie Illumina lors d'une pneumopathie humaine.*** Deux virus ont été mise en évidence : le virus respiratoire syncytial (RSV) et le cytomégalovirus (CMV) dont les génomes quasi complets ont été acquis. L'assemblage *de novo* (assembleur CLC Genomics) a identifié des contigs homologues chez ces deux virus. L'ensemble des *reads* non assemblés ont alors été localisés sur ces génomes de référence (mappeur CLC). Chaque graphe présente en abscisse les positions nucléotidiques du génome et en ordonnée la couverture du séquençage, c'est-à-dire le nombre de fois où chaque position est séquencée (source Pathoquest et Institut Pasteur du Cambodge [P. Buchy]).

débit considérable pour une durée de séquençage de l'ordre de quatre heures. Avec une bonne optimisation des outils de bio-informatique, obtenir une réponse en deux jours deviendra alors possible, pour un coût acceptable dans le cadre d'une utilisation hospitalière réservée à certains malades.

La sensibilité de certains *pipelines* (ce terme désignant l'ensemble des séquences allant du traitement de l'échantillon jusqu'à l'analyse bio-informatique des séquences, *voir Glossaire*) est au moins équivalente à celle de la PCR (*polymerase chain reaction*) ciblée, mais sans qu'il soit nécessaire de définir *a priori* des cibles. Nous avons utilisé récemment ce *pipeline* dans différents syndromes infectieux chez l'homme (encéphalites, pneumopathies), préalablement criblés négativement avec les outils classiques de PCR. À nouveau, le diagnostic par métagénomique nécessite d'assembler *de novo* de longs contigs au sein d'échantillons biologiques complexes *(Figure 4)*. La métagénomique met en évidence des agents pathogènes connus mais qui n'ont pas été identifiés par PCR, souvent en raison d'amorces ne couvrant pas l'étendue des séquences possibles. Elle représentera sans nul doute rapidement une alternative « moderne » aux techniques de culture maintenant abandonnées. Celles-ci avaient l'immense avantage d'être (pratiquement) sans *a priori* mais, malheureusement, elles ne détectaient que les micro-organismes cultivables. La métagénomique pérennisera l'avantage d'une détection large sans *a priori*, mais sans la conditionner à des capacités d'isolement.

Une crainte récurrente est que ce type d'approche ne trouve que des agents infectieux sans impact médical, et soit au final plus une source de confusion qu'une aide médicale. Tout d'abord, il est aisé de filtrer les résultats de manière à ce que ne soit délivrée au médecin que la liste des agents pathogènes présents. Ensuite, si l'on adopte une vision prospective, certaines associations de virus, bactéries et champignons révélées par la métagénomique devraient avoir une valeur diagnostique et orienter de manière significative le traitement. Il est certain que l'utilisation de ces outils non biaisés, communs à tous les types de microorganismes, va ouvrir un nouveau champ dans le diagnostic des maladies infectieuses.

Métagénomique virale et connaissance épidémiologique

Dans un autre champ, la métagénomique virale est en train de bousculer la connaissance épidémiologique des infections d'intérêt sanitaire, notamment en redéployant les territoires géographiques d'échantillonnage et en prenant mieux en compte les zones sauvages [28]. Le nombre

d'espèces séquencées par cette approche augmente en effet considérablement la quantité d'informations génétiques disponible. Les reconstructions phylogénétiques qui sont déduites de ces informations peuvent inférer avec davantage de fiabilité les liens épidémiologiques entre les isolats à l'échelle spatiale et chronologique [29], et permettre d'établir des chaînes de transmission inter-hôtes [30]. Certaines études ont permis d'appréhender des événements liés à un franchissement d'espèces et au développement d'une émergence. À titre d'exemple, une étude de séquençage profond des virus *Influenza* de type A a montré que le passage du réservoir canard vers des hôtes aviaires inhabituels (poulet, dinde) s'accompagne souvent de l'émergence de modifications génétiques des virus [31]. Dans cette étude, les auteurs montrent l'émergence de deux types de délétions dans le gène codant pour la neuraminidase qui confèrent probablement aux nouveaux génotypes viraux un avantage adaptatif chez des hôtes aviaires inhabituels, notamment chez la dinde [31]. Par ailleurs, depuis 2009, une nouvelle génération de travaux a émergé dans le domaine de la pathologie des plantes ; ces données permettent non seulement d'analyser le génome global d'un écosystème ou d'un organisme, mais aussi de relier directement les séquences des agents pathogènes à leur hôte et/ou à une position géographique [32]. Ces travaux *princeps* d'écogénomique ont révolutionné la vision de la distribution des phytovirus en révélant que près de 70 % des plantes analysées étaient « virosées » (infectées par un virus) [32]. Cette innovation récente, couplée au fait que les plantes sont immobiles et donc « ré-échantillonnables » dans le temps, devrait permettre de mieux comprendre la dynamique spatiotemporelle de la diversité phytovirale d'un agro-écosystème. Enfin, grâce à l'utilisation d'algorithmes probabilistes de plus en plus performants, les données générées par NGS constitueront une source d'information inestimable pour modéliser la dynamique des infections virales et prédire les émergences. Ces études devraient permettre de mieux rationaliser les approches plus coûteuses de collecte de données épidémiologiques sur le terrain.

Métagénomique virale et dynamique d'évolution des pathogènes

Les virus, en particulier ceux qui ont un intérêt médical, se caractérisent généralement par un nombre élevé d'individus par génération, des cycles de réplication très courts et un fort taux de mutations. L'accumulation des diversités qui en résulte, observable à l'échelle d'une vie humaine, fait des virus des modèles de choix pour étudier l'évolution du vivant. L'évolution des virus a par ailleurs un impact en virologie médicale, avec des conséquences sur la pathogénicité, le franchissement de la barrière d'espèce, l'échappement aux antiviraux et aux vaccins. Le niveau de diversité généré dépend du type de virus concerné et des mécanismes en jeu. Ainsi, les virus à ARN simple brin sont les plus « plastiques » : le taux de substitution/site/an est de 10^{-3} à 10^{-5}, soit 4 à 6 $\log_{10}$ au-dessus du taux de substitution généralement observé chez les organismes pluricellulaires [33]. Par ailleurs, les recombinaisons ou les échanges de segments participent à la génération de cette diversité. Pour appréhender la dynamique d'évolution des pathogènes et la dispersion des populations virales (quasi-espèces), les premières études reposaient sur des approches lourdes de clonage suivi du séquençage des clones par la technique de Sanger. Les NGS ou HTS permettent désormais d'explorer directement la diversité virale *in situ*, pour peu que certaines précautions d'ordre méthodologique soient prises pour éliminer les erreurs d'amplification, de séquençage et surtout d'assemblage des séquences qui peut produire des chimères artéfactuelles [34]. Dans les domaines humains mais aussi vétérinaires, les NGS se développent pour étudier la diversité virale générée aussi bien par substitution que par recombinaison [34]. Des travaux sur les cardiovirus humains ont ainsi mis en évidence la trace d'une recombinaison avec un theilovirus de rat, suggérant que l'ancêtre commun de ces virus proviendrait du rat [35]. La puissance de ces méthodes permet par ailleurs de caractériser la dynamique évolutive des virus chez un hôte ou entre différents hôtes [36]. La génération de la diversité peut également être explorée dans différents sites chez un même hôte et ce en fonction du développement de l'infection [37]. Les NGS permettent aussi de caractériser de façon dynamique l'émergence de mutants échappant aux antiviraux ou aux vaccins [38]. Au-delà de la connaissance qu'elles révèlent sur l'évolution d'un virus en particulier, les NGS contribueront certainement, dans un proche avenir, à élucider certaines étiologies complexes chez l'homme, l'animal ou la plante, pouvant résulter d'interactions subtiles (recombinaison, complémentation, etc.) entre variants d'une même espèce virale ou entre virus co-infectant une même cellule hôte.

Conclusions

Les travaux de métagénomique virale, ou plus largement le mouvement actuel d'acquisition massive de jeux de données NGS, sont confrontés à un double défi : premièrement, le traitement bio-informatique de ces millions voire milliards de séquences (tri, inventaire, stockage, accessibilité, etc.) ; deuxièmement, les interprétations biologique, écologique ou thérapeutique de ces nouveaux jeux de données. On peut cependant entrevoir que les progrès qui seront réalisés dans l'organisation et l'intégration de cette masse de données permettront de dépasser un niveau purement descriptif des résultats en leur donnant du sens dans le contexte des différents fronts de recherche et selon une démarche scientifique. Les données qualitatives actuellement obtenues seront enrichies par des données quantitatives robustes basées sur les réplications d'analyses dans le temps et dans l'espace, et soumises à des analyses statistiques rigoureuses. Nonobstant ces obstacles, la métagénomique a permis, au cours de cette dernière décennie, de réévaluer plusieurs concepts en écologie et en pathologie. Le rôle des virus dans un organisme ou dans un écosystème est par exemple un sujet à nouveau débattu, avec l'idée que des virus mutualistes ou bien

commensaux existent probablement [39]. Cette nouvelle typologie du rôle des virus, associée aux nouveaux ordres de grandeur de la diversité virale observée chez un hôte, nécessitera de mieux prendre en compte les interactions entre virus présents au sein des hôtes eucaryotes et, pour les virus pathogènes, d'actualiser les postulats de Koch [27]. La pathologie, dont l'étude s'est enrichie au début du XXIe siècle d'une approche développée initialement en écologie, devra sûrement accentuer son rapprochement avec cette discipline scientifique afin de mieux comprendre l'évolution et la dynamique des virus pathogènes ou mutualistes qui colonisent les humains, les animaux et les plantes. ◊

SUMMARY

Pathology and viral metagenomics, a recent history

Human, animal and plant viral diseases have greatly benefited from recent metagenomics developments. Viral metagenomics is a culture-independent approach used to investigate the complete viral genetic populations of a sample. During the last decade, metagenomics concepts and techniques that were first used by ecologists progressively spread into the scientific field of viral pathology. The sample, which was first for ecologists a fraction of ecosystem, became for pathologists an organism that hosts millions of microbes and viruses. This new approach, providing without *a priori* high resolution qualitative and quantitative data on the viral diversity, is now revolutionizing the way pathologists decipher viral diseases. This review describes the very last improvements of the high throughput next generation sequencing methods and discusses the applications of viral metagenomics in viral pathology, including discovery of novel viruses, viral surveillance and diagnostic, large-scale molecular epidemiology, and viral evolution. ◊

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Handelsman J, Rondon MR, Brady SF, *et al*. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem Biol* 1998 ; 5 : R245-9.
2. Schmidt TM, Delong EF, Pace NR. Analysis of a marine picoplankton community by 16S ribosomal-RNA gene cloning and sequencing. *J Bacteriol* 1991 ; 173 : 4371-8.
3. Pace NR, Stahl DA, Lane DJ, *et al*. Analyzing natural microbial populations by rRNA sequences. *ASM News* 1985 ; 51 : 4-12.
4. Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 1997 ; 276 : 734-40.
5. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and *in-situ* detection of individual microbial-cells without cultivation. *Microbiol Rev* 1995 ; 59 : 143-69.
6. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004 ; 304 : 66-74.
7. Allander T, Emerson SU, Engle RE, *et al*. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* 2001 ; 98 : 11609-14.
8. Breitbart M, Salamon P, Andresen B, *et al*. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 2002 ; 99 : 14250-5.
9. Rosario K, Breitbart M. Exploring the viral world through metagenomics. *Curr Opin Virol* 2011 ; 1 : 1-9.
10. Djikeng A, Kuzmickas R, Anderson NG, *et al*. Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* 2009 ; 4 : e38499.
11. Delong EF, Preston CM, Mincer T, *et al*. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 2006 ; 311 : 496-503.
12. Dupré J, O'Malley MA. Metagenomics and biological ontology. *Stud Hist Philos Biol Biomed Sci* 2007 ; 38 : 834-46.
13. Methe BA, Nelson KE, Pop M, *et al*. A framework for human microbiome research. *Nature* 2012 ; 486 : 215-21.
14. Faust K, Sathirapongsasuti JF, Izard J, *et al*. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* 2012 ; 8 : e1002606.
15. Dupré J. Emerging sciences and new conceptions of disease; or, beyond the monogenomic differentiated cell lineage. *Euro J Phil Sci* 20121 ; 1 : 119-31.
16. Lai YP, Di Nardo A, Nakatsuji T, *et al*. Commensal bacteria regulate Toll-like receptor 3-dependent inflammation after skin injury. *Nat Med* 2009 ; 15 : 1377-U4.
17. Feng HC, Shuda M, Chang Y, *et al*. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 2008 ; 319 : 1096-100.
18. Phan TG, Kapusinszky B, Wang CL, *et al*. The fecal viral flora of wild rodents. *PLoS Pathog* 2011 ; 7 : e1002218.
19. Sauvage V, Le Gouil M, Cheval J, *et al*. A member of a new *Picornaviridae* genus is shed in pig feces. *Int J Infect Diseases* 2012 ; 16 : E456.
20. Sauvage V, Foulongne V, Cheval J, *et al*. Human polyomavirus related to African green monkey lymphotropic polyomavirus. *Emerg Infect Dis* 2011 ; 17 : 1364-70.
21. Sauvage V, Cheval J, Foulongne V, *et al*. Identification of the first human *Gyrovirus*, a virus related to chicken anemia virus. *J Virol* 2011 ; 85 : 7948-50.
22. Willner D, Furlan M, Haynes M, *et al*. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One* 2009 ; 4 : e7370.
23. Foulongne V, Sauvage V, Hebert C, *et al*. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One* 2012 ; 7 : e38499.
24. Hoffmann B, Scheuch M, Hoper D, *et al*. Novel Orthobunyavirus in cattle, Europe, 2011. *Emerg Infect Dis* 2012 ; 18 : 469-72.
25. Honkavuori KS, Shivaprasad HL, Briese T, *et al*. Novel Picornavirus in turkey poults with hepatitis, California, USA. *Emerg Infect Dis* 2011 ; 17 : 480-7.
26. Lauck M, Hyeroba D, Tumukunde A, *et al*. Novel, divergent simian hemorrhagic fever viruses in a wild Ugandan red colobus monkey discovered using direct pyrosequencing. *PLoS One* 2011 ; 6 : e19056.
27. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2012 ; 2 : 63-7.
28. Van den Brand JMA, van Leeuwen M, Schapendonk CM, *et al*. Metagenomic analysis of the viral flora of pine marten and European badger feces. *J Virol* 2012 ; 86 : 2360-5.
29. Pybus OG, Suchard MA, Lemey P, *et al*. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci USA* 2012 ; 109 : 15066-71.
30. Escobar-Gutierrez A, Vazquez-Pichardo M, Cruz-Rivera M, *et al*. Identification of hepatitis C virus transmission using a Next-Generation Sequencing approach. *J Clin Microbiol* 2012 ; 50 : 1461-3.
31. Croville G, Soubies SM, Barbieri J, *et al*. Field monitoring of avian influenza viruses: whole-genome sequencing and tracking of neuraminidase evolution using 454 pyrosequencing. *J Clin Microbiol* 2012 ; 50 : 2881-7.
32. Roossinck MJ, Saha P, Wiley G, *et al*. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Mol Ecol* 2010 ; 19 : 81-8.
33. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 2008 ; 9 : 267-76.
34. Beerenwinkel N, Günthard HF, Roth V, *et al*. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol* 2012 ; 3 : 1-16.
35. Blinkova O, Kapoor A, Victoria J, *et al*. Cardioviruses are genetically diverse and cause common enteric infections in South Asian children. *J Virol* 2009 ; 83 : 4631-41.
36. Bull RA, Eden JS, Luciani F, *et al*. Contribution of intra- and interhost dynamics to Norovirus evolution. *J Virol* 2012 ; 86 : 3219-29.
37. Wright CF, Morelli MJ, Thebaud G, *et al*. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol* 2011 ; 85 : 2266-75.
38. Nishijima N, Marusawa H, Ueda Y, *et al*. Dynamics of hepatitis B virus quasispecies in association with nucleos(t)ide analogue treatment determined by ultra-deep sequencing. *PLoS One* 2012 ; 7 : e35052.
39. Roossinck MJ. The good viruses: viral mutualistic symbioses. *Nat Rev Microbiol* 2011 ; 9 : 99-108.
40. Jordan B. Une révolution longuement attendue. *Med Sci (Paris)* 2008 ; 24 : 869-73.
41. Kanitakis J. Carcinome à cellules de Merkel : première implication connue d'un polyomavirus dans la genèse d'un cancer humain. *Med Sci (Paris)* 2008 ; 24 : 570-1.

TIRÉS À PART

P. Roumagnac

3.2. Procédés utilisés en métagénomique virale

3.2.1. Différents acides nucléiques ciblés

La métagénomique virale est dans la majorité des cas basée sur la purification ciblée des acides nucléiques viraux. Trois types d'extraction des ARN/ADN viraux ont majoritairement été utilisés : l'extraction des acides nucléiques issus de la purification des particules virales (Blinkova *et al.*, 2009; Candresse *et al.*, 2014; Donaldson *et al.*, 2010; Melcher *et al.*, 2008; Victoria *et al.*, 2009), l'extraction des petits ARNs associés aux mécanismes de silencing chez les plantes (Candresse *et al.*, 2014; Molina *et al.*, 2012; Qi *et al.*, 2009; Wu *et al.*, 2010; Xu *et al.*, 2012) et l'extraction des doubles brin d'ARN viraux (Decker and Parker, 2014; Roossinck *et al.*, 2010). Par ailleurs, la technique de Rolling Circle Amplification (RCA) est couramment utilisée pour enrichir les échantillons en ADN simple brin circulaire (Johne *et al.*, 2009).

Cette étape d'extraction des acides nucléiques est cruciale car elle doit permettre d'éviter la surreprésentation des acides nucléiques non viraux (bactéries, champignons, plantes…). En effet, dans la plupart des échantillons, les acides nucléiques viraux représentent une faible proportion des acides nucléiques totaux (Blomstrom, 2011; Delwart, 2007).

3.2.1.1. Les particules virales

Une des propriétés des virus permettant de les distinguer des cellules procaryotes et eucaryotes est leur taille. Ainsi, une des techniques couramment utilisées pour éliminer les cellules « contaminantes » est une filtration à 0.2µm (Thurber *et al.*, 2009). Le dogme de la taille réduite des virus a été récemment renversé par la découverte de virus aussi grands que des bactéries (Philippe *et al.*, 2013; Raoult *et al.*, 2004) pour lesquels la technique de filtration est évidemment inadaptée. Par ailleurs, certaines techniques utilisent des ultracentrifugations densité-dépendantes combinées ou non à l'étape de filtration afin de pouvoir concentrer les particules virales (Thurber *et al.*, 2009). Par la suite, des traitements via des nucléases vont permettre de réduire la composition des échantillons en acides nucléiques hôtes ou autres (Allander *et al.*, 2001; Blomstrom, 2011). Les acides nucléiques viraux protégés par des capsides sont donc supposés rester indemnes suite à ces digestions enzymatiques. Toutefois, certains virions sont plus affectés que d'autres par ces traitements, et leurs acides nucléiques peuvent être dégradés (Blomstrom, 2011). Cependant, une étude visant à comparer différentes méthodes d'enrichissement en particules virales a démontré que les abondances d'acides nucléiques viraux résultantes étaient comparables (Hall *et al.*, 2014).

3.2.1.2. Les doubles-brins ARN (dsRNA)

La réplication de la majorité des phytovirus (à ARN) passe obligatoirement par une phase d'ARN double brin, ce qui en a fait une cible de choix pour détecter l'infection phytovirale. Une méthode d'extraction basée sur l'affinité des dsRNA avec la cellulose

CF-11 a été mise au point bien avant les travaux de métagénomique afin de détecter des infections virales au sein de végétaux (Dodds *et al.*, 1984). De plus, ces structures de dsRNA sont plus stables que les ssRNA, ce qui permet d'éviter les pertes lors du processus d'extraction. Cependant cette technique ne va pas permettre de détecter les virus à ADN. De plus il a été reporté que les végétaux contenant des phénols, du latex ou d'autres métabolites secondaires étaient difficiles à extraire (Roossinck *et al.*, 2010).

3.2.1.3. Les petits ARNs issus des mécanismes de silencing

Nous avons déjà évoqué dans la section 2.1.1.1 que le mécanisme de RNA silencing chez les plantes, en réponse à une infection virale, va mener à la production de petits ARN de 21 à 24 nucléotides. Il a donc été entrepris dans diverses études de viser ces petits ARN pour détecter des virus à partir d'échantillons de végétaux (Kreuze *et al.*, 2009). Ainsi, une extraction d'ARN total est entreprise, une fois les ARN extraits, ils sont soumis à une migration sur gel d'acrylamide et les brins d'ARN ayant une taille située entre 20 et 30 nucléotides sont extraits et ligués à des adaptateurs permettant une amplification ciblée (Kreuze *et al.*, 2009). Ainsi ils vont pouvoir être directement séquencés et ne seront pas soumis à l'amplification séquence-indépendante décrite ci-après.

3.2.2. Amplifications « séquences-indépendantes » des acides nucléiques viraux

Contrairement à d'autres groupes d'organismes, les virus n'ont pas de gène en commun permettant de faire une phylogénie de tous les virus et qui pourrait être utilisé comme cible pour la métagénomique virale. Pour y remédier, et permettre la détection de virus inconnus, une étape d'amplification séquence-indépendante est nécessaire (Blomstrom, 2011) . Les méthodes les plus fréquemment utilisées sont : la Sequence-Independant Single-Primer Amplification (SISPA) (Reyes and Kim, 1991), la random-PCR (rPCR) (Froussard, 1993) et l'amplification via la polymérase du phage Phi29 (RCA) évoquée précédemment (Johne *et al.*, 2009). Afin de traiter plusieurs échantillons lors d'une même réaction de séquençage, on peut ajouter un identifiant moléculaire (MID pour molecular identifier, qui est souvent appelé « tag » ou « barcode ») à chaque échantillon, ce qui permet de regrouper les échantillons (multiplexage). L'adjonction de ces tags va permettre lors du traitement de données issues du séquençage, de réassigner chaque séquence générée à son échantillon d'origine (démultiplexage). La rPCR va permettre d'intégrer ces MIDs. Cependant, il a été montré que certains MIDs ne sont parfois pas fonctionnels dans le sens où ils ne permettent pas d'avoir une amplification (Roossinck *et al.,* 2010). De plus, ces amplifications peuvent générer un biais concernant la composition en bases azotées des amplicons ou leur représentativité. En effet, certaines séquences peuvent être préférentiellement amplifiées (Dong *et al.*, 2011; Hamady and Knight, 2009; Yilmaz *et al.*, 2010), et les séquences rares noyées dans un océan de séquences majoritaires peuvent ne pas être amplifiées (Sharpton, 2014).

Par ailleurs, lorsque l'on traite plusieurs échantillons à la fois, que ce soit lors de la concentration des particules virales ou lors de l'amplification de leurs séquences, il existe un risque inhérent de contamination (Degnan and Ochman, 2012). Cette contamination amène à la génération de faux positifs (Rosseel *et al.*, 2014) qui vont être difficile à identifier et à enlever lors des analyses bioinformatiques (Kunin *et al.*, 2008). *In fine*, les contaminations peuvent amener à une vision biaisée de la diversité de la communauté. Il existe cependant des logiciels permettant d'identifier et d'enlever les contaminants (Schmieder and Edwards, 2011).

3.2.3. Nouvelles techniques de séquençage (NGS)

Il faut bien différencier métagénomique de NGS. La métagénomique est un concept, les NGS sont un outil souvent utilisé en métagénomique. Pour preuve, les premières études de métagénomique n'utilisaient pas des NGS, elles utilisaient le séquençage Sanger (séquençage de première génération). Il existe plusieurs nouvelles technologies de séquençage à ce jour (Tableau SB.2).

Sequencing platform	Amplification method	Sequencing chemistry	Read length (bp)	Sequencing Speed/h	Maximum Output Per run	Accuracy (%)	M^1 I^2 D^3
454 (Roche)	Emulsion PCR	Pyrosequencing	400–700	13 Mbp	700 Mbp	99.9	0.10, 0.3, 0.02 [23]
Illumina (Illumina)	Bridge PCR	Reversible terminators	100–300	25 Mbp	600 Gbp	99.9	0.12, 0.004, 0.006 [23]
SOLiD (Life Technologies)	Emulsion PCR	Ligation	75–85	21–28 Mbp	80–360 Gbp	99.9	Error is higher than Illumina [24]
PacBio (Pacific Biosciences)	No amplification Single molecule real-time (or SMRT)	Fluorescently labeled nucleotides	4, 000–5,000	50–115 Mbp	200 Mb–1 Gbp	95	1, 2, 12 [25]
Helicos (Helicos Biosciences)	No amplification Single molecule	Reversible terminators	25–55	83 Mbp	35 Gbp	97	Error is in the range of few percent but higher than 454 and Illumina and biased toward InDels [24]
Ion Torrent (Life Technologies)	Emulsion PCR	Detection of released H	100–400	25 Mb–16 Gbp	100 Mb–64 Gbp	99	M, 0.06, I + D 1.38 [26]
Nanopore (Oxford Technologies)	No amplification Single molecule		Very long reads up to 50 kbp	150 Mbp	Tens of Gbp	96	

M^1 = Mismatch bases; I^2 =Insertion; D^3 = Deletion.

Tableau SB.2 : Les nouvelles technologies de séquençage (NGS) les plus couramment utilisées. D'après Barba *et al.*, 2014.

On parle de seconde génération de séquençage pour les méthodes qui vont séquencer les acides nucléiques via un processus d'amplification alors que l'on parle de troisième génération de séquençage lorsqu'il n'y a pas d'amplification liée au processus de séquençage. Les facteurs qui doivent être pris en compte dans la sélection d'une plateforme de séquençage NGS sont : la taille du génome à séquencer, sa complexité (par exemple son pourcentage en GC), le taux d'erreur de la technologie, la profondeur et la couverture attendue. Parmi les NGS de seconde génération, le pyroséquençage (454) a longtemps fournit les reads les plus longs alors que l'Illumina HiSeq2500 a le meilleur rendement et est le moins cher. En revanche la fiabilité est meilleure pour SOLiD (Liu *et al.*, 2012b). Cependant, une étude de comparaison entre le 454, Illumina HiSeq et Ion Torrent PGM a montré que le choix de la technologie de séquençage avait un impact négligeable sur les résultats en termes de représentativité des séquences dans les métagénomes (Solonenko *et al.*, 2013). À ce jour, Roche a arrêté la commercialisation du 454. Dans les années futures on s'attend à ce que les NGS de 3e génération accroissent

leur capacité de séquençage et leur rapidité avec une réduction du coût. Un génome humain devrait être séquencé pour 1000 dollars ce qui est presque 10 fois moins cher que le prix actuel (Barba *et al.*, 2014).

3.2.4. Traitements bioinformatiques des données NGS

Les données brutes massives produites par les NGS nécessitent une batterie de traitements. Des outils ont été mis en place afin de permettre *in fine* de déterminer quels virus sont présents dans les échantillons et d'aller jusqu'à prédire les fonctions de leurs gènes. Ces outils évoluent constamment pour s'adapter aux progrès des NGS, et notamment à l'augmentation vertigineuse des données qu'ils produisent. Cette augmentation nécessite également de faire évoluer les capacités informatiques (Droge and McHardy, 2012; Hunter *et al.*, 2012; Stapleton, 2014).

3.2.4.1. Nettoyage de données

Comme indiqué précédemment, les NGS génèrent des erreurs lors du séquençage. Les jeux de données bruts issus du séquençage vont donc dans un premier temps être soumis à des filtres et des contrôles de qualité permettant de ne garder que les séquences qui se situent à certains seuils de « propreté ». De plus, si les séquences sont issues d'un multiplexage, c'est à cette étape qu'on va pouvoir réattribuer chaque séquence à son échantillon d'origine (démultiplexage).

3.2.4.2. Répartition en classes ou « binning »

De manière générale, la répartition en classes (dite « binning » en anglais) va permettre de regrouper des séquences sur la base d'un pourcentage de similarité ou de caractéristiques intrinsèques telles que leur contenu en GC. Ces séquences peuvent par exemples être regroupées en OTUs (Operational Taxonomic Units). L'avantage de cette méthode est qu'elle permet de réduire la complexité des données qui vont être soumises aux analyses ultérieures. Divers logiciels, dont MEGAN (Huson *et al.*, 2007), MG-RAST (Meyer *et al.*, 2008) ou encore QIIME (Caporaso *et al.*, 2010) permettent de faire de la répartition en classes.

Lors des analyses basées sur les reads d'un jeu de donnée de métagénomique, seule une sous-partie de ce jeu de donnée va être assignée à des fonctions ou à des taxas. La qualité des assignations peut varier en fonction (i) des bases de données de référence utilisées pour les comparaisons et (ii) de la longueur des reads (Scholz *et al.*, 2012; Sibley *et al.*, 2012). De plus, il a été démontré que des reads courts n'étaient pas appropriés à la caractérisation des communautés microbiennes en métagénomique (Wommack *et al.*, 2008). Pour pallier à ce problème on peut réaliser un assemblage à partir des reads.

3.2.4.3. Assemblage des reads

L'assemblage permet d'unifier des reads colinéaires provenant théoriquement du même génome en une seule et même séquence contigüe (appelée contig) avec le risque,

cependant, de créer des chimères. L'assemblage génère des séquences plus longues que celles du jeu de données de base issues des NGS et permet parfois d'obtenir des génomes entiers (Candresse *et al.*, 2014; Culley *et al.*, 2007; Kreuze *et al.*, 2009; Tse *et al.*, 2012). Par ailleurs, il est possible d'utiliser la répartition en classe préalablement à l'assemblage (Scholz *et al.*, 2012; Sharpton, 2014). L'avantage associé à l'assemblage est qu'il est plus probable d'associer des séquences longues à un Blast. Bien qu'il existe une large gamme d'algorithmes permettant d'effectuer des assemblages (Scholz *et al.*, 2012), peu ont été adaptés à l'analyse de métagénomes ; parmi ces derniers on compte entre autres MetaVelvet (Namiki *et al.*, 2012), Meta-IDBA (Peng *et al.*, 2011) et Genovo (Laserson *et al.*, 2011).

Comme signalé ci-dessus, il peut y avoir création de chimères lors de l'assemblage. Par exemple, quand deux reads sont assemblés dans un même contig, il n'est pas possible de savoir si ces deux reads proviennent du même organisme ou de deux organismes différents (Charuvaka and Rangwala, 2011; Mavromatis *et al.,* 2007; Mende *et al.*, 2012; Pignatelli and Moya, 2011; Scholz *et al.*, 2012; Vazquez-Castellanos *et al.*, 2014). Ainsi, l'assemblage peut mener *in fine* à une sous-estimation de la diversité des organismes. De plus, l'assemblage est limité aux taxa les plus abondants de la communauté (reads ayant de fortes couvertures) et cette opération nécessite de larges capacités informatiques et prend un temps considérable (Blomstrom, 2011; Melcher *et al.*, 2014; Scholz *et al.*, 2012; Sharpton, 2014; Wooley and Ye, 2010). Cependant, l'amélioration des NGS et l'avènement du séquençage de 3^{e} génération vont permettre de produire des reads plus long, ce qui limitera la création de chimères au moment de l'assemblage et diminuera les temps de traitements informatiques (Melcher *et al.*, 2014; Scholz *et al.*, 2012).

3.3.Etude des métagénomes

Une large gamme d'outils informatiques permet de centraliser l'analyse des métagénomes allant du nettoyage du jeu de données à son analyse (Tableau SB.3). Concernant l'analyse de ces métagénomes, une étude de la biodiversité des communautés échantillonnées est inévitable, mais on peut aller plus loin qu'un simple inventaire de biodiversité. En effet il est possible de comparer les métagénomes, et selon le plan d'échantillonnage, il est possible d'analyser l'association de certains paramètres écologiques avec la composition des communautés. Cependant, une large quantité de données est nécessaire pour obtenir des résultats probants. En effet, il a été démontré qu'une couverture insuffisante limitait le pouvoir des analyses statistiques appliquées aux métagénomes (Rodriguez and Konstantinidis, 2014) .

Resource	Methods	Citation	Web link
AmphoraNet	Marker gene analysis: phylogeny	Kerepesi et al. (2014)	http://pitgroup.org/amphoranet/
CAMERA	Various: taxonomic and functional annotation, comparative analyses	Sun et al. (2011)	http://camera.calit2.net/
Comet	Functional annotation, comparative analyses	Lingner et al. (2011)	http://comet.gobics.de/
LEfSe (Galaxy)	Comparative analyses	Segata et al. (2011)	http://huttenhower.sph.harvard.edu/galaxy/
IMG/M	Various: taxonomic and functional annotation, comparative analyses	Markowitz et al. (2014)	https://img.jgi.doe.gov/m/
MG-RAST	Various: taxonomic and functional annotation, comparative analyses	Meyer et al. (2008)	http://metagenomics.anl.gov/
MALINA	Various: taxonomic and functional annotation, comparative analyses	Tyakht et al. (2012)	http://malina.metagenome.ru/
METAGENassist	Various: taxonomic annotation, comparative analyses	Arndt et al. (2012)	http://www.metagenassist.ca/
MetaPhlAn (Galaxy)	Marker gene analysis: similarity	Segata et al. (2012)	http://huttenhower.sph.harvard.edu/galaxy/
NBC	Binning: compositional classification	Rosen et al. (2011)	http://nbc.ece.drexel.edu
Orphelia	Gene prediction	Hoff et al. (2009)	http://orphelia.gobics.de/
Phylopithia webserver	Binning: compositional classification	Patil et al. (2012)	http://phylopythias.cs.uni-duesseldorf.de/
Real time metagenomics	Functional annotation	Edwards et al. (2012)	http://edwards.sdsu.edu/rtmg/
WebCARMA	Binning: sequence similarity	Gerlach et al. (2009)	http://webcarma.cebitec.uni-bielefeld.de/
WebMGA	Various: taxonomic and functional annotation	Wu et al. (2011)	http://weizhong-lab.ucsd.edu/metagenomic-analysis/

Tableau SB.3 : Différentes ressources pour l'analyse de métagénomes. D'après Sharpton, 2014.

3.3.1. Composition taxonomique

Bien avant l'avènement de la métagénomique, les écologistes précisaient déjà que déterminer la distribution et l'abondance des espèces d'un habitat ou d'un écosystème était le point de départ pour les études macro-écologiques des communautés (biodiversité, composition, structure, fonctions, associations) (Begon *et al.*, 1990; Hubbell, 2001).

A partir du binning ou des assemblages, des comparaisons de séquence utilisant le logiciel Blast (Altschul *et al.*, 1997) peuvent être réalisées. Ce logiciel permet de comparer une séquence, nucléique ou protéique, dite requête, à une banque de séquences, nucléiques ou protéiques via la réalisation d'alignements de séquences et de calculs des similarités des séquences alignées. Blast compare les séquences nucléiques sur les 2 brins, c'est-à-dire la séquence étudiée (brin +) et la séquence complémentaire inversée (brin -). Pour comparer une séquence nucléique à une séquence protéique, Blast traduit la séquence nucléique en générant toutes les séquences protéiques possibles, c'est-à-dire 6 séquences différentes (3 à partir du brin + et 3 à partir du brin -). Ces travaux permettent de révéler la composition taxonomique des échantillons. Une multitude de logiciels permettant ces affiliations sont disponibles mais ils sont généralement dédiés à l'analyse de communautés bactériennes. Cependant, des logiciels tels que GAAS (Angly *et al.*, 2009), ProViDE (Ghosh *et al.*, 2011), ou Metavir (Roux *et al.*, 2014) permettent maintenant de déterminer les compositions taxonomiques de communautés virales issues de métagénomique. A partir de ces affiliations taxonomiques on peut alors inférer la structure des communautés issues de la

métagénomique en calculant des indices de diversité α (voir section 3.3.2) et en matérialisant la proportion/répartition de chaque virotype au sein du métagénome.

Cependant, des travaux de recherche ont démontré la présence d'erreurs systématiques dans des métagénomes générés par le 454 qui amènent à une surestimation de l'abondance en taxon et en gènes de 11% à 35% (Gomez-Alvarez *et al.*, 2009). Cette surestimation serait due à la création artificielle de réplicats lors du pyroséquençage. De plus l'abondance des pathogènes est faiblement corrélée avec leur concentration dans l'échantillon (Yang *et al.*, 2011). Ces études suggèrent donc de prendre des précautions, lors de l'analyse des estimations d'abondances des travaux de métagénomique.

3.3.2. Analyses de la diversité dans les métagénomes

3.3.2.1. Méthodes

Avant de procéder à une quelconque mesure de diversité, les séquences virales (reads ou contigs) sont regroupés suivant un certain pourcentage d'identité ; on appelle cela un cluster ou encore un virotype. On connaît ainsi le nombre de séquences affiliées à chaque cluster. Dans le domaine de la bactériologie, ces clusters sont effectués avec un seuil de similarités de 97% et sont appelés OTUs.

De nombreuses études de métagénomique virale présentent des diagrammes « en camembert » qui reflètent les diversités virales. Ces diagrammes sont construits à partir des Blasts et du nombre de reads ou contigs qui les représentent à un certain niveau taxonomique. Ainsi on peut avoir une représentation de l'abondance de séquences par taxa. Toutefois, des outils informatiques utilisant des méthodes d'analyse de la diversité virale ont été mis en place. Parmi ces outils, METAVIR (Roux *et al.*, 2011; Roux *et al.,* 2014) permet la création de diagrammes d'abondance évoqués précédemment et permet de générer des courbes de raréfaction. Le logiciel QIIME (Caporaso *et al.*, 2010) dédié aux communautés bactériennes et fongiques a également été récemment utilisé dans une étude de métagénomique virale et a permis d'illustrer la diversité alpha par des courbe de raréfaction et de la calculer sur la base des indices de Chao1 et de Shannon-Wiener (Perez-Brocal *et al.*, 2013). L'outil le plus couramment utilisé est PHACCS (Angly *et al.*, 2005). En estimant la richesse et l'équitabilité des taxa viraux en se basant sur un spectre de contigs, PHACCS va pouvoir générer des courbes de rang-abondance et calculer des indices de diversité (Indice de Shannon-Wiener). Ainsi, PHACCS a permis d'évaluer la diversité virale dans diverses communautés telles que des échantillons d'eau ou encore de fèces (Angly *et al.*, 2006; Breitbart *et al.,* 2008; Minot *et al.*, 2011; Reyes *et al.*, 2010; Roux, 2012). Récemment, un nouveau programme d'estimation de la diversité des communautés microbiennes issues de NGS a été mis en place. Ce logiciel nommé CatchAll (Bunge, 2011; Bunge *et al.*, 2012) évalue la diversité en se basant non pas sur des courbes de rang-abondance mais sur des comptages de fréquences de séquences permettant ainsi de leur appliquer des tests statistiques. Des comparaisons de résultats issus de PHACCS et de CatchAll ont montré que PHACCS sous-

estime probablement les richesses spécifiques (Allen *et al.*, 2013). D'autres outils tels que UniFrac (Lozupone *et al.*, 2006) ou MEGAN (Huson *et al.*, 2007) représentent la diversité sous la forme de phylogénies avec au bout de chaque branche un cercle de taille variable représentant le nombre de séquences appartenant à chaque taxa. MEGAN a notamment été utilisé dans la fameuse étude de Craig Venter en 2004 visant à analyser la diversité des communautés microbiennes de la mer des Sargasses (Venter *et al.*, 2004).

Toutes ses méthodes ont été implémentées afin de faire des mesures de la diversité α qui est bien souvent représentée par la richesse spécifique, et aucune n'a été à ce jour désignée comme la méthode standard (Bunge *et al.*, 2014).En ce qui concerne la diversité β, les écologistes microbiens utilisant les NGS ont opté pour des analyses statistiques (ANOVA, analyse en composante principale…) permettant de comparer des communautés plutôt que pour des calculs d'indices (Bunge *et al.*, 2014; Perez-Brocal *et al.*, 2013; Roux, 2012). Cependant des logiciels tels que METAVIR (Roux *et al.*, 2011; Roux *et al.*, 2014) permettent une visualisation sous forme d'arbre de la proximité de différents métagénomes en termes de contenu en séquences.

3.3.2.2. Limitations

Il existe plusieurs limitations à l'usage des approches de métagénomique pour étudier et estimer la diversité virale présente au sein de volumes d'échantillonnage délimités (organe, organisme, populations, communautés etc.). Les biais liés à l'échantillonnage, le traitement des échantillons, et l'analyse des données peuvent induire des biais dans la représentativité des espèces de l'échantillon (Bunge *et al.*, 2014).

Par ailleurs, comme nous le concluons dans l'article de revue publié dans Médecine & Science, les données qualitatives actuellement obtenues devront être enrichies par des données quantitatives robustes basées sur les réplications d'analyses dans le temps et dans l'espace, et soumises à des analyses statistiques rigoureuses.

3.3.3. Comparaison des métagénomes et étude de l'influence de paramètres biotiques et abiotiques sur les métagénomes

La comparaison des métagénomes est essentielle si l'on veut comprendre l'influence de l'environnement sur la structuration et l'évolution des communautés. Différentes approches de comparaison ont ainsi été mises au point, elles peuvent être basées sur différents paramètres des métagénomes, e.g. leur composition en oligonucléotides (dinucléotides, tétranucléotides, etc.), leur taille, leur composition taxonomique ou encore leur contenu fonctionnel (Wooley and Ye, 2010). Plusieurs logiciels ou pipelines ont été dédiées à ces analyses (Angly *et al.*, 2005; Blankenberg *et al.*, 2010; Goll *et al.*, 2010; Huson *et al.*, 2007; Lozupone *et al.*, 2006; Meyer *et al.*, 2008).

Les analyses multivariées telles que l'Analyse en Composante Principale (ACP) et les calculs de distances entre métagénomes telles que le « Non Metric-Multi Dimensional

Scaling » (NM-MDS) sont typiquement utilisés afin de visualiser les données et révéler quels sont les facteurs qui affectent les données (Brulc *et al.*, 2009; Willner *et al.*, 2009a; Willner *et al.*, 2009b). La méthode du NM-MDS a notamment permis de démontrer une séparation génétique des communautés virales en fonction de la salinité de l'eau (Roux, 2013) (Figure SB.24). Par ailleurs, l'ACP a permis de démontrer que les communautés de phages de patients atteint de cystite fibreuse étaient liés à l'environnement respiratoire interne, les communautés de phages des individus malades sont similaires alors que celles des individus sains sont elles distinctes les unes des autres (Willner *et al.*, 2009a). De plus, le « Permutation Tail Probability (PTP) test » utilise des arbres phylogénétiques pour déterminer si des groupes taxonomiques sont préférentiellement associés à un environnement donné. Par exemple, il a été démontré que les communautés marines de phages sont phylogénétiquement proches les unes des autres quels que soient leur lieu d'échantillonnage (Breitbart *et al.*, 2002).

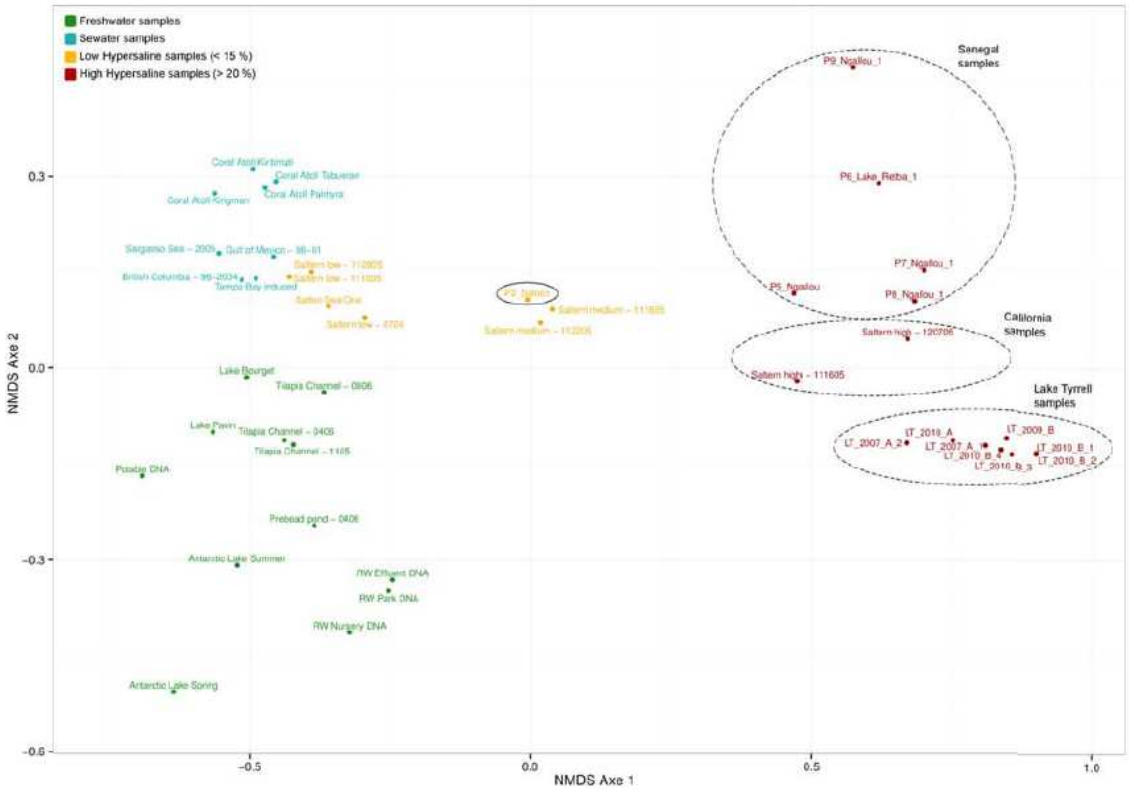


Figure SB.24 : Comparaison globale de viromes aquatiques. Des Blasts ont été effectués à partir de sous-échantillons de chaque virome (50 000 séquences de 100pb), les résultats des Blasts ont été utilisés dans une comparaison via NM-MDS basée sur leur similarité génétique. Les viromes sont colorés en fonction du degré de salinité de l'échantillon. Les cercles en pointillés entourent 3 groupes d'échantillons hypersalins. Le cercle plein entoure l'échantillon qui a une salinité intermédiaire. D'après Roux, 2013.

Les comparaisons de métagénomes basées sur leur contenu nucléotidique (GC ou oligonucléotides) permettent de s'affranchir des affiliations taxonomiques, ce qui permet de tenir compte des séquences n'ayant aucune similarité avec des séquences connues. En effet, suite à la comparaison de 41 viromes d'origines diverses (eaux, animaux, coraux…), il s'est avéré qu'il existait une signature génomique propre à chaque type d'environnement au niveau des fréquences des motifs dinucléotidiques des

métagénomes qui expliquerait 80% de la variance entre biomes (Willner *et al.*, 2009b). Ainsi une hypothèse a été émise selon laquelle la sélection imposée par l'environnement serait à l'origine de ces signatures (Willner *et al.*, 2009b). En utilisant le même type d'approche, il a été démontré que le type d'écosystème est un élément structurant des communautés en termes de fonctions (Dinsdale *et al.*, 2008). Par ailleurs, la comparaison de métagénomes sur la base de leur contenu en GC, a déjà été appliquée dans le but de comparer des communautés microbiennes issues d'échantillons marins et édaphiques (Yooseph *et al.*, 2007).

La comparaison des métagénomes sur la base de leurs affiliations taxonomiques, prenant souvent en compte des indices de diversité (cf. sections 1.6.3 et 3.3.2), ont également permis de mettre en évidence l'importance de facteurs biotiques et abiotiques dans la structuration des communautés virales (Minot *et al.,* 2011; Perez-Brocal *et al.*, 2013; Reyes *et al.*, 2010; Roux, 2012). Des études sur les viromes humains ont notamment permis de démontrer (grâce aux indices de diversité) que des jumelles homozygotes possédaient des flores virales spécifiques différentes alors que leur flore bactérienne étaient similaires (Reyes *et al.,* 2010) ou encore qu'une convergence dans le type de régime alimentaire amenait à une convergence de la composition taxonomique des viromes (Minot *et al.*, 2011).

Cependant il existe un problème majeur lorsque l'on veut comparer des métagénomes issus d'études différentes. En effet, les méthodes utilisées pour générer les données ne sont souvent pas les mêmes, ce qui peut donc induire un biais dans les résultats issus de telles analyses (Wooley and Ye, 2010). Par conséquent, la nécessité de protocoles expérimentaux et analytiques communs a été soulignée dans le cadre de l'application de la métagénomique en tant que méthode de diagnostic (Dunne *et al.*, 2012).

3.3.4. Études menées en virologie végétale utilisant la métagénomique

Bien que les études de métagénomique aient débuté dans les années 1990 (Schmidt *et al.*, 1991), leur application en virologie végétale n'a réellement débuté qu'en 2009 (Barba *et al.*, 2014). Très récemment, Barba et ses collègues ont réalisé une revue d'articles qui balaie 64 études utilisant les NGS en virologie végétale (Barba *et al.,* 2014). Cette revue complète notre article de revue présenté au début de ce chapitre. Même si les 64 études ne relèvent pas toutes du domaine de la métagénomique, pour celles qui en font partie, il en ressort diverses applications (détection, identification, découverte, transcriptomique, écologie, épidémiologie) que ce soit au niveau individuel, de quelques individus ou d'une communauté entière. Nous présenterons ci-dessous les résultats récents de métagénomique phytovirale sous un angle « écologique » - de l'individu-hôte aux communautés végétales afin de compléter les résultats présentés dans la revue Médecine et Science qui proposait alors un angle plus « pathologique ».

3.3.4.1. Études des métagénomes viraux à l'échelle d'une plante

Comme décrit précédemment, différentes méthodes ont été appliquées en métagénomique phytovirale afin de détecter et/ou caractériser les viromes présents au sein de végétaux. Différentes types de traitement d'échantillons ont été effectués tels que des extractions à partir de VLPs (Virus Like Particles) ou des extractions ciblant les ARN viraux (totaux ou dsRNA) ou encore ciblant les virus circulaires à ADN (Barba *et al.*, 2014). Par ailleurs, de nombreuses études de métagénomique phytovirale se sont focalisées sur les petits ARNs générés lors du mécanisme de silencing induit par les infections virales dans les plantes (Giampetruzzi *et al.,* 2012; Kashif *et al.*, 2012; Kreuze *et al.,* 2009; Loconsole *et al.*, 2012; Pallett *et al.*, 2010). Une étude pionnière dans ce domaine a par exemple mené à la détection de deux virus connus et à la découverte de deux nouveaux virus dans des plants de patate douce (Kreuze *et al.,* 2009). D'autres études à partir de métagénomes viraux issus des petits ARNs ont par ailleurs permis de mettre en évidence des mécanismes liés au silencing incluant la biogénèse des siRNAs (Donaire *et al.,* 2009; Lin *et al.*, 2010; Navarro *et al.,* 2009; Pantaleo *et al.*, 2010; Qi *et al.*, 2009; Ruiz-Ruiz *et al.*, 2011; Szittya *et al.*, 2010; Xu *et al.*, 2012). Par ailleurs, une étude réalisée à partir d'ARN totaux a permis d'estimer l'intensité de la sélection et de la dérive génétique sur des variants du *Potato virus Y* (Fabre *et al.,* 2012). La métagénomique a également été mise à profit dans le cadre du diagnostic de plants de quarantaine, permettant ainsi de mettre en évidence une infection virale masquée par un autre virus dans des plants de canne à sucre (Candresse *et al.*, 2014) ou encore de caractériser une nouvelle espèce de marafivirus à partir de plants servant à la culture bioénergétique (Agindotan *et al.*, 2010).

Ces études ont ainsi permis de mettre en évidence la présence de virus connus au sein de végétaux ainsi que des infections multiples. Ces travaux se sont aussi traduits par la découvertes de nouvelles espèces virales sur des hôtes cultivés (la majorité d'entre elles ciblant la vigne) mais aussi sur hôtes sauvages (Barba *et al.*, 2014). Par ailleurs, une étude originale de métagénomique a mené à la reconstruction d'un génome à ARN de *Barley stripe mosaic vir*us issue d'une graine d'orge datant de 750 ans, amenant ainsi à un questionnement quant à l'histoire évolutive des phytovirus et à leur émergence (Smith *et al.*, 2014).

3.3.4.2. Études des métagénomes viraux à l'échelle d'un groupe de plantes

Quelques études ont été menées à partir de groupes (« pools ») de quelques échantillons d'individus de la même espèce ou d'espèces différentes (Coetzee *et al.*, 2010; Wyant, 2012; Wylie *et al.*, 2011). Par exemple, une étude de métagénomique menée à partir d'un pool de 17 plantes sauvages (dont 14 plantes australiennes indigènes) a permis la détection et l'identification de 12 virus connus et de 4 nouveaux virus. Des amorces ont été dessinées à partir des séquences issues des travaux de métagénomique et ont permis un retour sur les échantillons végétaux d'origine permettant ainsi d'associer chaque virus à sa plante hôte (Wylie *et al.,* 2013). Une autre

étude menée sur un pool de 44 échantillons de feuilles issues d'un vignoble a permis l'identification de 4 virus associés à la vigne, mais aucune étude n'a été entreprise afin d'identifier quels individus étaient infectés (Coetzee *et al.*, 2010). Toutefois, toutes ces études, que ce soit au niveau individuel ou d'un pool d'individus, ont été menées à partir de plantes symptomatiques.

3.3.4.3. Études des métagénomes à l'échelle des communautés végétales

Les études évoquées précédemment ont surtout été menées au niveau individuel ou sur une faible quantité d'individus. À ce jour, seulement deux études à grande échelle font état de la biodiversité phytovirale présente au sein d'une communauté de plantes et ce sans *a priori* (les plantes ont été récoltées sans se soucier du fait qu'elles présentaient des symptômes ou non). Une étude d'écogénomique basée sur les dsRNA viraux présents au sein de centaines d'échantillons de plantes sauvages de l'aire nationale de conservation de Guacanaste au Costa-Rica a permis l'identification de 11 familles de phytovirus infectant les plantes (70% de plantes positives à la présence de virus – i.e. reads/contigs) et à la découverte de milliers de nouveaux virus de plantes (non publié). Ces travaux ont aussi montré que les infections multiples sont probablement un phénomène commun dans les espaces naturels. L'originalité de ces travaux d'écogénomique vient du fait que chaque échantillon végétal a été géo-référencé et associé à un MID ; ainsi il a été possible d'assigner chaque séquence virale à son hôte et à sa localisation géographique (Roossinck *et al.*, 2010). Une seconde étude visant les particules virales de 95 plantes sauvages (52 espèces) récoltées dans la Tallgrass Prairie Preserve en Oklahoma, a montré que 25% des plantes étaient infectées (10 espèces sur les 52) (Muthukumar *et al.*, 2009).

Ce type d'étude effectuée sur les communautés de virus associées aux communautés végétales est primordial pour la compréhension de l'incidence et de la propagation des virus, fournissant ainsi des indices sur les émergences mais aussi sur le rôle des virus dans les écosystèmes. Par ailleurs, en plus d'apporter des indices quant à l'écologie des phytovirus, la métagénomique phytovirale apparaît aujourd'hui comme un outil permettant d'évaluer le risque pour les cultures représenté par les phytovirus (MacDiarmid *et al.*, 2013; Roossinck, 2012b).

3.3.4.4. Autres ressources pour analyser la diversité phytovirale

Des virus de plantes ont été détectés dans d'autres sources que des végétaux. En effet, des phytovirus ont pu être détectés à partir d'insectes vecteurs (Ng *et al.*, 2011a), de sol (Kim *et al.*, 2008), d'organes humains (Walker, 2010), de fèces (Zhang *et al.*, 2006) ou d'échantillons d'eaux de diverses origines (Ng *et al.*, 2012; Rosario *et al.*, 2009; Roux, 2012). Ces études montrent que la stabilité de certains phytovirus en dehors de leurs hôtes peut être importante, ce qui pourrait jouer un rôle - encore largement inexploré - dans la conservation et la dissémination de certains virus des plantes. Ces résultats ne semblent cependant pas être extrapolables à tous les virus des plantes car les virus

considérés comme étant les moins stables n'ont pas été retrouvé dans ces échantillons (Roossinck, 2012b).

3.4. Limitations conceptuelles de la métagénomique et solutions envisagées

3.4.1. Limitations conceptuelles liées à la métagénomique

En plus d'être liée à des limitations techniques citées précédemment, la métagénomique virale est également liée à des problèmes d'ordre conceptuel.

Premièrement, comme évoqué dans la partie 1 de cette introduction, le concept d'espèce virale est débattu, et il l'est d'autant plus quand on veut l'appliquer à la métagénomique. En effet, lors des affiliations taxonomiques, il est nécessaire d'imposer des seuils permettant de délimiter les taxas. La définition du seuil va alors être cruciale si l'on veut interpréter les variations entre génomes en fonction des facteurs biotiques et abiotiques (Ward *et al.,* 2008). En ce qui concerne les espèces bactériennes il est communément admis que des séquences ayant un pourcentage de similarité supérieur à 98% au niveau du gène codant pour le 16S rARN font partie de la même espèce (Thompson *et al.*, 2013), cependant aucun seuil commun à toutes les espèces virales n'a été (et ne peut être) déterminé. En effet, pour chaque genre viral, les membres de l'ICTV ont déterminé un seuil propre à la démarcation des espèces qui lui appartiennent.

Deuxièmement, il faut prendre des précautions vis-à-vis des affiliations taxonomiques déduites des analyses Blast. En effet, si une séquence générée par NGS est similaire à une séquence de la base de données, ce n'est pas une condition suffisante pour affirmer qu'elle provient d'un virus du même taxon. Par exemple, si une séquence issue d'un échantillon X a pour séquence la plus proche celle d'un geminivirus, deux classements taxonomiques sont possibles : soit X fait partie intégrante des *Geminiviridae*, soit X est un « outgroup » de cette famille. Il serait donc plus approprié d'associer à cette séquence l'affiliation de « *Geminiviridae*-like » plutôt que celle de « *Geminiviridae* ».

Troisièmement, la plus grande question posée vis-à-vis des virus détectés via la métagénomique est : infectent-ils réellement les individus analysés et s'y répliquent-ils ? En effet, lors de l'identification de séquences virales dans un échantillon analysé via la métagénomique, on détecte une présence (basée sur des reads ou des contigs) et non pas une infection (Canuti and van der Hoek, 2014). On va ainsi préférer attribuer le terme VLPs (Virus Like Particles) à ces séquences virales car on ne connaît pas la proportion de virus compétant pour la réplication (Minot *et al.*, 2011). De plus les séquences virales détectées peuvent en réalité être issues de génomes viraux intégrés à celui de leur hôte (Canuti and van der Hoek, 2014). Dans le but de démontrer l'infection il va falloir isoler le virus, le transmettre et le ré-isoler afin de vérifier les postulats de Koch. En effet, une étude de métagénomique ayant détecté un cucumovirus sur une plante symptomatique a été complétée par une vérification des postulats de Koch

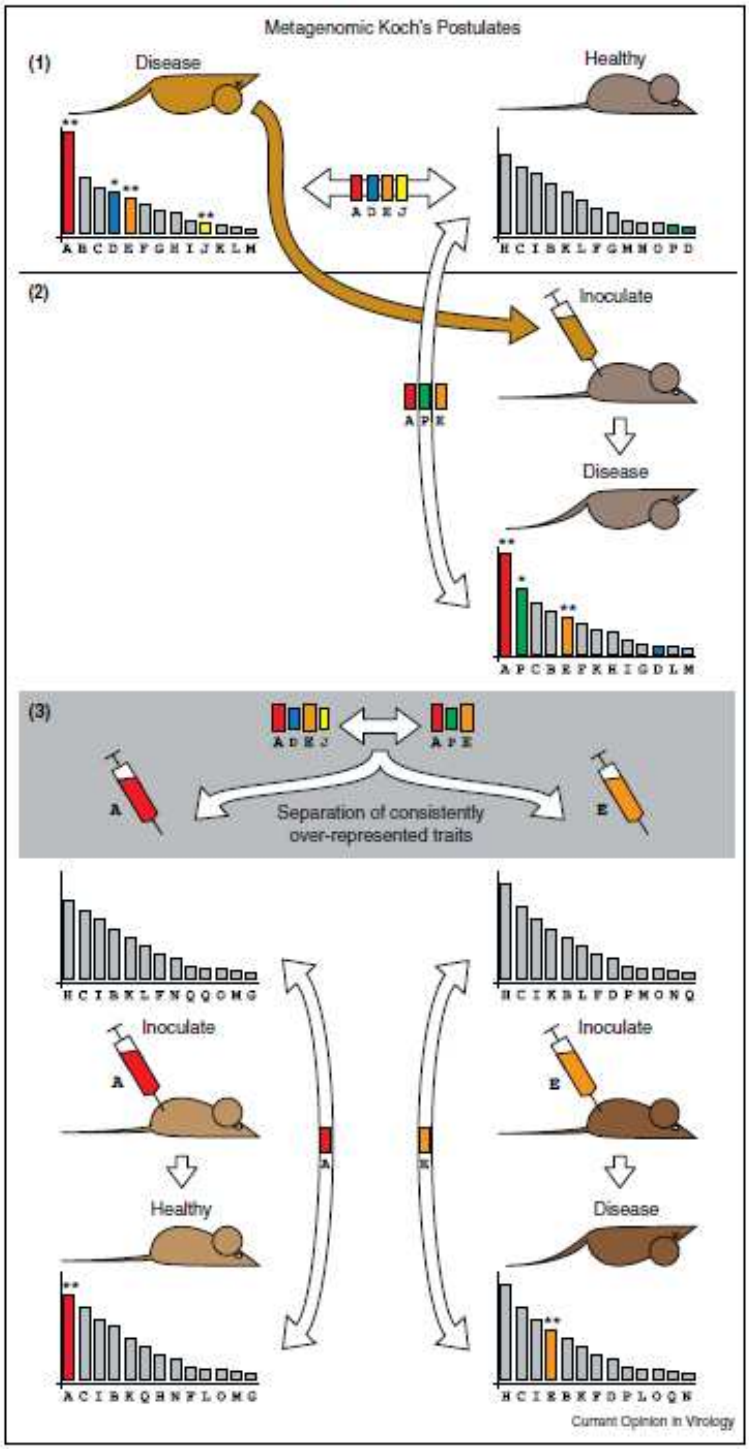


Figure SB.25: Le postulat de Koch en métagénomique. La comparaison entre un animal sain et un animal malade indique une différence significative au sein des librairies de métagénomique (illustrées par les histogrammes montrant l'abondance relative des reads). Pour vérifier le postulat de Koch en métagénomique, il faut (1) que les traits métagénomiques du sujet malade soient significativement différents du sujet sain (par exemple les traits A, D, E et J, soient présents chez le sujet malade et non chez le sain, (2) que l'inoculation d'échantillons provenant de l'animal malade à un animal sain induise la maladie (la comparaison des métagénomes avant et après inoculation doit suggérer l'acquisition ou l'accroissement de nouveaux traits métagénomiques (A, E, et P), les nouveaux traits peuvent alors être purifiés via des méthodes telles que la dilution en série, (3) que l'inoculation des traits purifiés dans un animal sain induise la maladie. D'après Mokili *et al.*, 2012.

démontrant ainsi que le virus identifié n'était pas à l'origine des symptômes (Adams *et al.*, 2009). Cependant, ces postulats sont, en pratique, difficiles à tester (Canuti and van der Hoek, 2014). John L. Mokili propose un « metagenomic Koch postulate » qui se base sur des marqueurs moléculaires utilisés en tant que traits métagénomiques et il conseille ainsi de comparer ces traits sur des patients sains et malades (Figure SB.25) (Mokili *et al.*, 2012). Certains auteurs se demandent notamment si les séquences virales issues de métagénomique devraient avoir leur propre système de classification et de nomenclature (MacDiarmid *et al.*, 2013; Melcher *et al.*, 2014).

Quatrièmement il paraît frustrant d'obtenir une majorité de séquences sans identité (matière noire, Figure SB.26) et de ne pas les traiter. Comme évoqué précédemment, une des solutions peut être de comparer les signatures oligonucléotidiques qui pourraient être propres aux virus d'un environnement donné (Dinsdale *et al.*, 2008; Willner *et al.*, 2009b). De plus, la circularisation des contigs ou la structure des cadres de lecture (ORFs) pourraient également être des indices de présence virale (Mokili *et al.*, 2012). Par exemple, une étude se focalisant sur les séquences pouvant se circulariser (virus à ssDNA circulaire putatifs) a permis de mettre en évidence 129 virotypes issus de 608 génomes. Ces virotypes n'ayant aucune similarité au niveau des bases de données, il a été entrepris d'analyser leurs ORFs, ceci a alors montré un codage potentiel pour des ORFs Rep et CP. Ainsi de nouvelles familles virales distantes de celles déjà connues ont été mises en évidence (Labonte and Suttle, 2013).

(a)

Reference	Sample type	Method	Cut off	Unknown
Breitbart et al., 2002	Seawater	TBLASTX	E value =0.001	65%
Breitbart et al., 2003	Human feces	TBLASTX	E value = 0.001	59%
Breitbart et al., 2004	Marine sediment	TBLASTX	E value = 0.001	75%
Zhang et al., 2006	Human feces	TBLASTX	E value = 0.001	91%
Benchet al., 2007	Virioplankton	TBLASTX	E value = 0.00	61%
Breitbart et al., 2008	Infant feces	TBLASTX	E value = 0.001	66%
Rosarion et al, 2009	Potable water	TBLASTX	E value = 0.001	44%
Rosarion et al, 2009	Reclaimed water (DNA)	TBLASTX	E value = 0.001	70%
Rosarion et al, 2009	Reclaimed water (RNA)	TBLASTX	E value = 0.001	57%
Desnues et al., 2008	Stromatolite	TBLASTX	E value = 0.01	99%

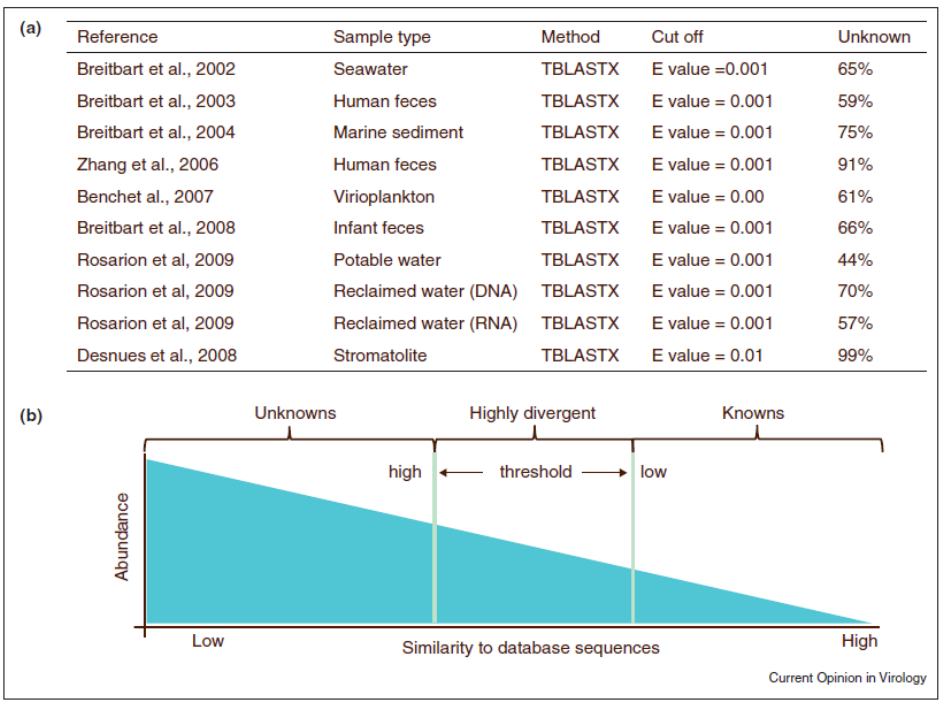


Figure SB.26: La matière noire ou séquences inconnues: les séquences qui n'ont pas d'homologues dans GenBank. (a) Proportion de séquences inconnues dans des études de métagénomes issus de divers environnements. (b) Diagramme illustrant l'abondance des séquences inconnues et des séquences connues dans l'environnement. La distinction entre connues et inconnues dépend du seuil utilisé. D'après Mokili *et al.*, 2012.

Compte tenu des limitations techniques et conceptuelles actuelles liées aux travaux de métagénomique, une étape de vérification associée à un retour aux techniques classiques de biologie moléculaire et de culture est alors recommandé pour vérifier la validité des jeux de données issus de travaux de métagénomique (Bibby, 2013). En effet, un retour à la technique de PCR peut montrer un désaccord avec les séquences générées via la métagénomique (Yang *et al.*, 2011). Par ailleurs, lorsque l'on découvre un nouveau virus, il va être important de mettre en place des études complémentaires permettant d'évaluer le risque qui lui est associé (Canuti and van der Hoek, 2014).

3.4.2. Retour aux techniques classiques

Afin de compléter les résultats issus de métagénomique et de mieux les interpréter, il est nécessaire de retourner à des études et concepts de virologie plus classique. A titre d'exemple, l'étude concernant la diversité des phytovirus dans la « Tallgrass prairie preserve » en Oklahoma a conduit à la découverte de nombreux virus infectant des plantes sauvages (Muthukumar *et al.*, 2009; Roossinck *et al.*, 2010). Les équipes de recherche d'Ulrich Melcher et de Marylin Roossinck ont ainsi entrepris de pousser cette étude plus loin en se focalisant sur un virus détecté via la métagénomique sur la plante sauvage *Asclepias viridis* ayant une séquence proche de tymovirus. À partir de ce résultat ils ont pu isoler, cloner et caractériser ce virus (génome et phylogénie), étudier sa gamme d'hôte, ses symptômes et sa virulence et de manière plus large la distribution spatio-temporelle de ce virus a pu être évaluée dans la zone d'échantillonnage (Min *et al.*, 2012). Une autre équipe a développé un test de rt-PCR à partir de résultats de métagénomique pour analyser la distribution d'un nouveau virus (Marafivirus) au sein de plants de vigne et de cicadelles collectés et piégés dans les vignobles californiens (Al Rwahnih *et al.*, 2009).

4. Problématique-Présentation du sujet

La synthèse bibliographique présentée ici révèle la multiplicité et la complexité des interactions plantes-virus qui se déroulent à différentes échelles (de l'organisme au paysage). Elle montre aussi que les études menées en phytovirologie se sont principalement focalisées sur des virus issus du milieu cultivé, et que peu d'études concernant le milieu sauvage ont été effectuées. Ce manque de connaissance des milieux naturels et des interactions plante-virus qui s'y déroulent représente un écueil dans notre compréhension de l'écologie et de l'évolution des phytovirus sur le long terme. Cette quasi-absence de connaissance ne permet en outre pas de totalement comprendre, modéliser et prédire les processus micro- et/ou macro-évolutifs qui se mettent en place à l'échelle de l'agro-écosystème. Ainsi il semble encore très difficile de pouvoir comprendre la nature des interactions plantes-virus dans les écosystèmes et ainsi pouvoir modéliser et prédire l'évolution des « maladies » en milieu naturel et l'émergence, la résurgence ou la persistance de maladies en milieu agricole.

Plusieurs questions restent donc sans réponse et nous avons décidé d'essayer d'y répondre dans le cadre de ce travail de thèse :

• Le milieu sauvage est-il un réservoir de biodiversité phytovirale ? En d'autres termes, la diversité végétale a-t'elle un impact sur la diversité virale ?

• La diversité végétale influence t'elle les prévalences virales ?

• Y a-t'il des patrons de distribution spatio-temporelle des phytovirus dans l'agro-écosystème ?

• Quels paramètres écologiques permettent d'expliquer ces distributions ?

La synthèse bibliographique a par ailleurs souligné le potentiel de la métagénomique en tant qu'outil d'étude des communautés virales et plus précisément du virome des plantes. Elle a aussi montré que l'agro-écosystème, à l'interface entre milieu naturel et milieu cultivé, est probablement un lieu d'étude de choix pour répondre aux questions formulées ci-dessus. Ainsi le chapitre I de cette thèse est consacré à une étude de géo-métagénomique phytovirale au sein de deux agro-écosystèmes. Les objectifs principaux de ce chapitre sont d'essayer de répondre aux trois questions définies précédemment.

Nous avons cependant vu que plusieurs biais sont encore associés aux travaux de métagénomique, nous suggérant qu'il serait probablement judicieux de mener des études plus approfondies de virologie « classique » sur quelques modèles viraux d'intérêt révélés par les approches de géo-métagénomique. Ainsi le chapitre II de cette thèse se focalise sur l'étude d'un nouveau genre viral découvert grâce à nos travaux de géo-métagénomique : le genre Capulavirus (*Geminiviridae*). Ce chapitre II est divisé en plusieurs parties qui ont pour objectif de répondre aux questions suivantes :

• Quelles sont les caractéristiques propres aux Capulavirus en termes d'organisation génomique, de gamme d'hôte et de vection ?

• Quelle est leur prévalence, leur diversité et leur répartition géographique ?

• Où se placent-t-ils dans la phylogénie des *Geminiviridae* et quelles informations apportent-ils quant à la compréhension de l'histoire évolutive de cette famille virale ?

Chapitre I : Etude de l'influence de l'agriculture sur la diversité, la prévalence et la distribution spatio-temporelle des phytovirus au sein de deux agro-écosystèmes à l'aide de la géo-métagénomique.

1. Introduction

La synthèse bibliographique présentée en début du manuscrit a permis de souligner que les activités humaines ont eu pour effet une simplification des écosystèmes qui s'est traduite par une réduction de la diversité végétale. Ces changements ont entrainé une modification des paramètres écologiques qui influencent les communautés d'organismes présents au sein de ces écosystèmes. Ce constat a amené les scientifiques à émettre et à tester des hypothèses sur l'effet de cette réduction de diversité végétale sur les émergences phytovirales. Une des hypothèses considère que la diversité des organismes hôtes et non-hôtes aurait un effet sur les prévalences des organismes pathogènes (Keesing *et al.*, 2010; Keesing *et al.*, 2006; Pagan *et al.*, 2012; Rodelo-Urrego *et al.*, 2013). Par ailleurs, quelques études décrivent le milieu sauvage comme réservoir de diversité phytovirale sans pour autant disposer de points de comparaison (Muthukumar *et al.*, 2009; Roossinck, 2012b; Roossinck *et al.*, 2010).

La synthèse bibliographique a par ailleurs montré que les hotspots de biodiversité en contact avec des zones agricoles semblent être des espaces appropriés à l'étude du lien entre la diversité végétale et la diversité phytovirale. Ces interfaces semblent favorables à l'évaluation de l'impact de l'Homme sur la distribution spatio-temporelle des phytovirus et sur leur écologie. À ce jour, aucune étude de la diversité des phytovirus à l'échelle d'un agro-écosystème n'a été réalisée. De plus, les études faisant état de la diversité globale des phytovirus dans un écosystème naturel se sont limitées à l'inventaire des virus présents au sein de l'écosystème (Muthukumar *et al.*, 2009; Roossinck *et al.*, 2010) sans pouvoir aborder la distribution spatio-temporelle des phytovirus et la relier à des paramètres écologiques.

Pour ce premier chapitre de thèse, nous avons donc décidé d'étudier la diversité, la prévalence et la distribution spatio-temporelle des phytovirus associés à deux agro-écosystèmes : le fynbos (Afrique du Sud) et la Camargue (France). Nous allons ainsi nous attacher à (i) estimer si le milieu sauvage (non-anthropisé) est un réservoir de biodiversité phytovirale et si la diversité végétale a un impact sur la diversité et la prévalence virale ; (ii) analyser si il existe des patrons de distribution spatio-temporelle des phytovirus dans l'agro-écosystème ; et (iii) étudier les paramètres écologiques qui permettent d'expliquer les distributions et prévalences virales.

2. Matériels et Méthodes

2.1. Sites d'étude

Les campagnes d'échantillonnage réalisées lors de cette étude ont été menées sur deux sites appartenant à deux écosystèmes distincts soumis à un climat de type méditerranéen. Le premier écosystème correspond au fynbos de la région floristique du Cap en Afrique du Sud. Cette région est considérée comme un hotspot de biodiversité, en effet, elle abriterait environ 8200 espèces végétales, dont 6200 plantes endémiques (Myers *et al.*, 2000). Le deuxième écosystème correspond à la Camargue située dans le

bassin méditerranéen en France. L'ensemble du bassin méditerranéen est lui aussi considéré comme un hotspot de biodiversité avec une estimation de 25 000 espèces de plantes, dont 11700 plantes endémiques (Myers *et al.*, 2000). Ces deux régions, qui sont aussi caractérisées par de larges zones de cultures intensives (vigne, céréales etc.), possèdent *de facto* une multitude d'interfaces agro-écologiques. Ces deux zones se distinguent cependant par le fait que l'Afrique du Sud présente des interfaces incluant des zones naturelles (fynbos) peu ou non perturbées par l'Homme alors que la Camargue présente des zones « semi »-naturelles, le plus souvent résilientes, ayant été durablement perturbées par l'action de l'Homme (endiguement du Rhône depuis le Moyen-âge, pâturage extensif ou intensif, urbanisation etc.) avant d'être partiellement protégées (Réserve Naturelle de Camargue, Réserve de la Tour du Valat, Parc Naturel de Camargue) depuis quelques décennies.

La campagne d'échantillonnage concernant le fynbos a été réalisée au niveau d'une interface agro-écologique de la région du Cap (Bufflesfontein Game and Nature Reserve et cultures avoisinantes) et la campagne d'échantillonnage concernant la Camargue a été réalisée au niveau de la station biologique de « La Tour du Valat » et des cultures avoisinantes (Figure 1.1).

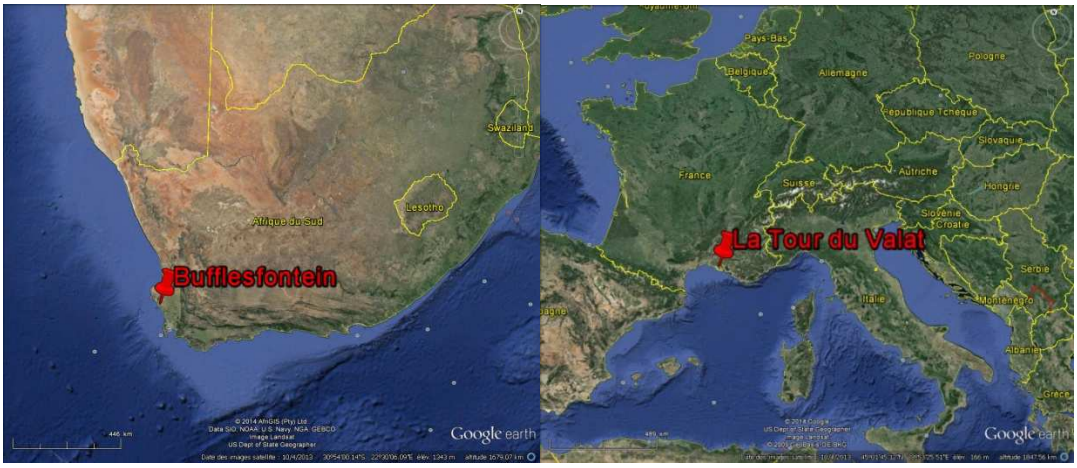


Figure 1.1: Localisation géographique des sites d'échantillonnage en Afrique du Sud (Bufflesfontein) et en France (La Tour du Valat).

2.2. Échantillonnage

Sur chacun des deux sites d'étude nous avons positionné une grille d'échantillonnage virtuelle de 20,25 km^2 (4,5 x 4,5km) au niveau d'une interface entre des zones agricoles et non-agricoles (Figures 1.2 et 1.3). Chaque grille est composée de 100 points d'échantillonnage géo-référencés ; ces points sont espacés de 500 m les uns des autres. Au niveau de chaque point d'échantillonnage (aire de 1 m^2), nous avons collecté environ 10 g des parties aériennes des espèces végétales majoritaires. Chaque échantillon est ainsi constitué d'un à plusieurs individus végétaux de la même espèce. L'échantillonnage a été effectué « à l'aveugle », c'est-à-dire sans tenir compte de l'expression de possibles symptômes associés à des infections virales. Ainsi, nous avons

collecté en moyenne 5 à 6 espèces différentes par point en Camargue et 7 à 8 espèces différentes par point dans le fynbos. Ce type d'échantillonnage a été effectué 4 fois :

(1) période de mai-juin 2010 en Camargue (C2010), (2) mai-juin 2012 en Camargue (C2012), (3) septembre-octobre 2010 dans le fynbos (F2010), et (4) septembre-octobre 2012 dans le fynbos (F2012). Nous avons respectivement collecté 495, 630, 743 et 713 échantillons végétaux pour C2010, C2012, F2010 et F2012. Les échantillons ont été stockés individuellement à 4°C durant les périodes d'échantillonnage et les transports, puis, une fois au laboratoire, à -80°C.

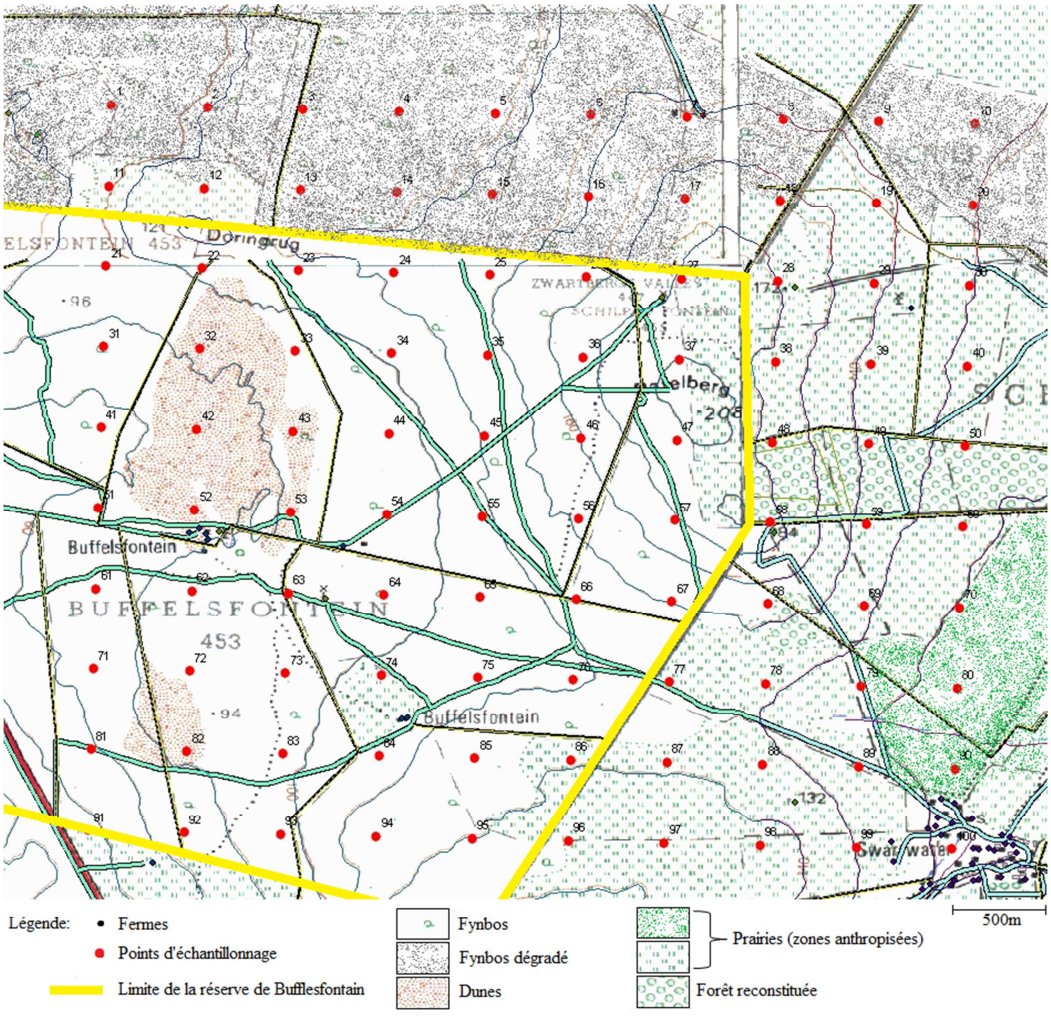


Figure 1.2 : Disposition de la grille d'échantillonnage en Afrique du Sud.

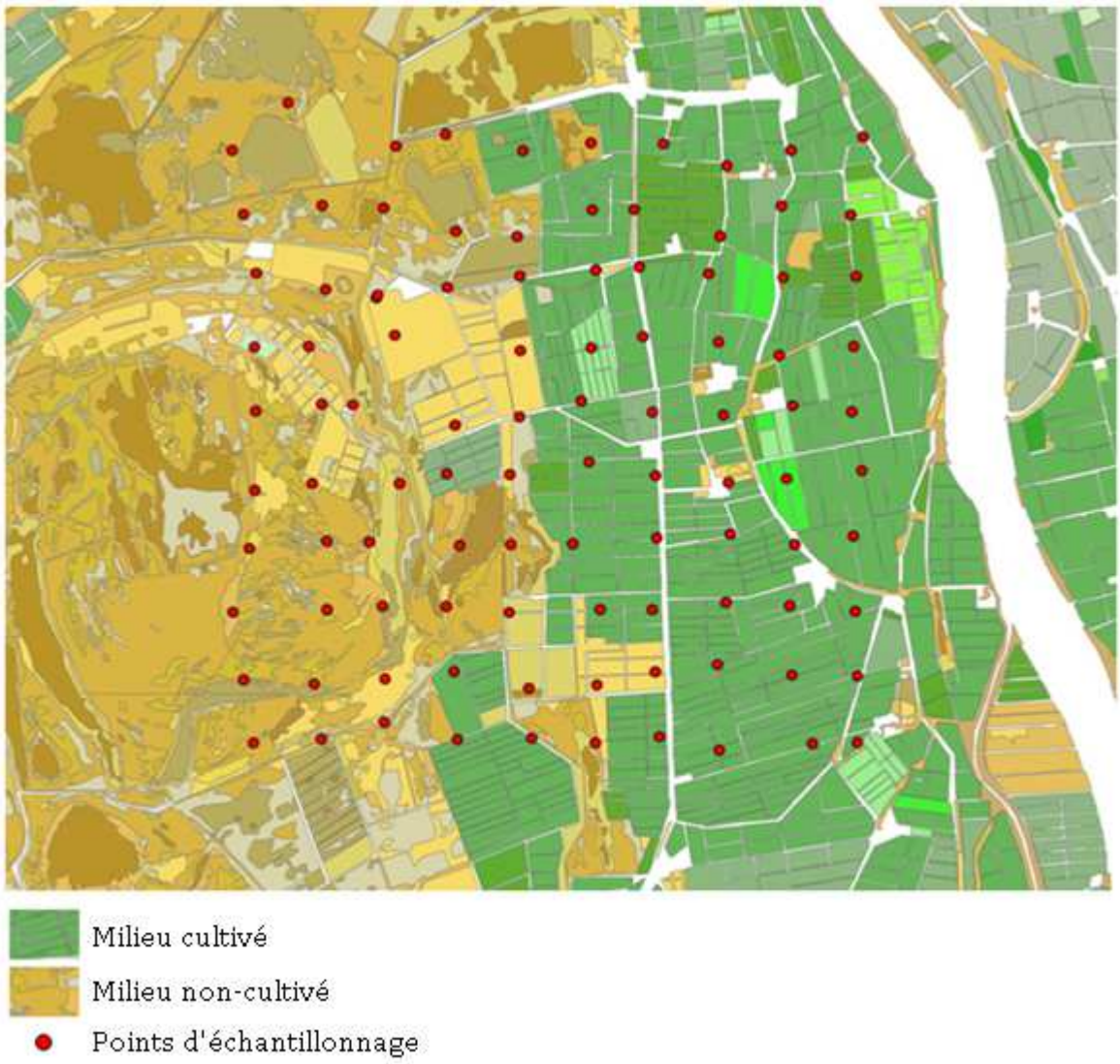


Figure 1.3 : Disposition de la grille d'échantillonnage en Carmargue.

2.3.Identification des échantillons végétaux

L'identification des échantillons végétaux au niveau du genre voire de l'espèce a été effectuée par des botanistes de la Tour du Valat et du Jardin botanique de Kirstenbosch sur la base de caractères morphologiques. Les échantillons qui n'ont pas pu être identifiés de cette manière ont été soumis à une extraction d'ADN puis à un typage génétique (barcoding) via l'amplification et le séquençage du gène chloroplastique *rbc*L. Le barcoding effectué à partir du gène *rbc*L est reconnu pour balayer un large spectre de plantes et permet de déterminer le genre de l'échantillon mais ne permet pas de descendre au niveau de l'espèce (Hollingsworth *et al.*, 2009). Avant d'effectuer les manipulations de barcoding, les échantillons végétaux ont été soumis à une extraction d'ADN à l'aide du kit DNeasy 96 Plant Kit (Qiagen) en suivant le protocole du fournisseur. À partir des extraits d'ADN nous avons procédé à une amplification du gène chloroplastique *rbc*L à l'aide des amorces rbcL*a*-F (5'-ATGTCACCACAAACAGAGACTAAAGC-3') (Levin *et al.*, 2003) et rbcL*a*-R (5'-GTAAAATCAAGTCCACCRCG-3') (Kress *et al.*, 2009). La PCR a été effectuée avec le GoTaq

Green Master Mix (Promega) : pour chaque échantillon le milieu réactionnel est constitué de 1µL d'extrait d'ADN (entre 15 et 250ng/µL), 12.5µL de GoTaq Green Master Mix, 0.5µL de chaque amorce (à 10µM chacune) et 10.5µL d'eau. Le cycle PCR est le suivant : un cycle de 2min à 95°C, 35 cycles de 1min à 94°C, 1min à 55°C, et 1min à 72°C, puis une extension finale de 5min à 72°C. Les produits d'amplification ont été migrés sur gel d'agarose à 1% durant 20min à 100V. Les produits PCR de taille attendue ont été séquencés par la société Beckman Coulter Genomics via la technique de Sanger. Les séquences ont été comparées à la banque de séquence GenBank par BlastN afin de déterminer le genre végétal des échantillons correspondants.

2.4. Typologie des habitats

A chaque point d'échantillonnage a été attribué un score associé au type d'habitat auquel il correspond (Figure 1.4). Ces scores reflètent un gradient concernant les niveaux de perturbation du milieu par l'activité agricole.

Le score 0 correspond aux habitats naturels. Ce sont des milieux non ou très faiblement perturbés ou influencés par l'Homme, c'est-à-dire dont les processus écologiques sont quasiment vierges d'influence humaine.

Le score 1 a été attribué aux habitats naturels « peu perturbés ». Ce sont des milieux qui ont été perturbés ou influencés par l'Homme lors des dernières décennies (e.g. pâturage ovin ou bovin extensif en Camargue ou dans la région du Cap) mais dans lesquels aucune plante cultivée n'a été introduite (roselières, sansouires, milieu dunaire ou prés salés en Camargue ; zones à fynbos – Renosterveld ou Sandveld au Cap).

Le score 2 a été attribué aux habitats semi-naturels. Ce sont des milieux qui ont été influencés ou perturbés par l'Homme, notamment lors de l'introduction de plantes cultivées (e.g. forêts, friches liées à des rotations longues ou à de la déprise agricole). Toutefois, la majeure partie des processus écologiques de ces habitats ne résulte pas d'une gestion humaine directe ni n'est fortement perturbée par cette dernière.

Le score 3 a été attribué aux cultures diversifiées et aux prairies menées de façon extensive. Ce sont des milieux d'origine agricole.

Le score 4 a été attribué aux habitats artificialisés. Ce sont des milieux contrôlés et conduits de manière intensive par l'Homme. Cette gestion agricole a vraisemblablement conduit à la perte de la majorité des processus écologiques encore partiellement actifs dans les deux catégories précédentes. Il est à noter que ces zones artificialisées accueillent encore quelques espèces sauvages capables de survivre à ces environnements extrêmement modifiés (adventices).

À partir de cette typologie des habitats, nous avons choisi d'établir deux grands types de milieux : le milieu non-cultivé comprenant les habitats de scores 0 et 1 (habitats naturels et naturels perturbés sans présence de cultures), et le milieu cultivé

comprenant les habitats scorés de 2 à 4 (friches, forêts, cultures diversifiées, prairies, et habitats artificialisés).

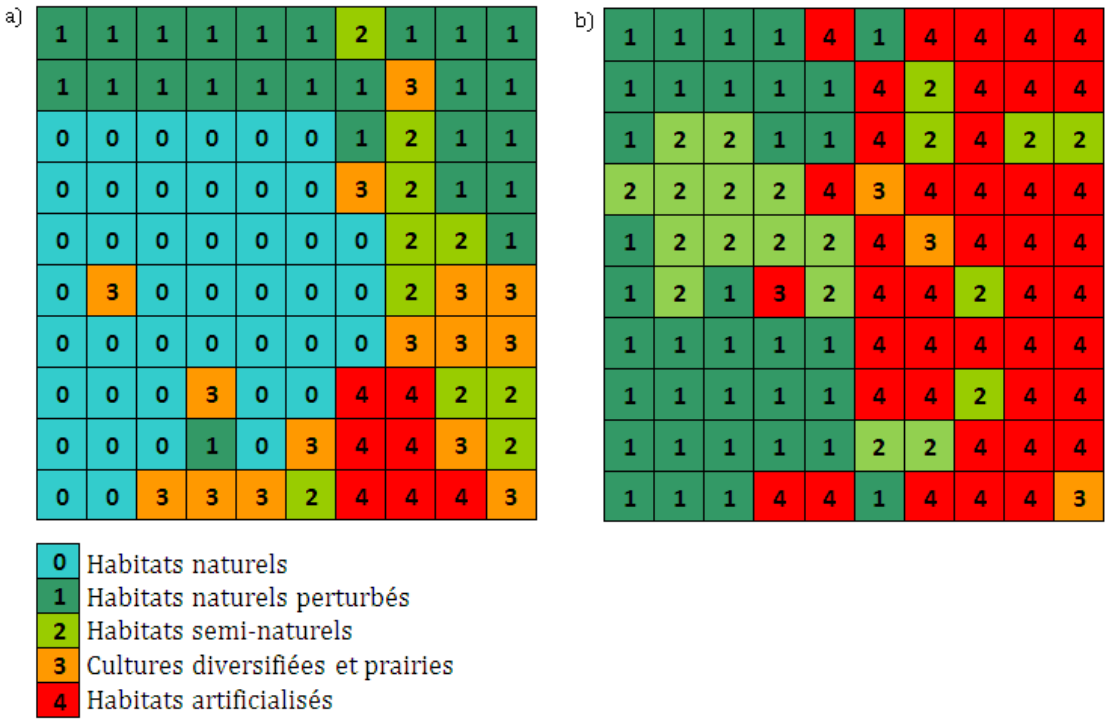


Figure 1.4 : Scores associés au type d'habitat de chaque point d'échantillonnage pour le fynbos (a) et la Camargue (b).

2.5.Extraction des acides nucléiques viraux à partir de particules virales, amplifications et séquençage

Chaque échantillon végétal a été traité individuellement et soumis à une extraction des acides nucléiques viraux suite à une semi-purification de particules virales. Les acides nucléiques viraux extraits ont alors été soumis à une série d'amplifications. La technique de semi-purification a été adoptée car elle permet de traiter un grand nombre d'échantillons en un espace de temps relativement court et de cibler tous les types d'acides nucléiques viraux. Le protocole suivi est décrit dans la publication publiée dans la revue PLoS ONE à laquelle j'ai participé : Appearances can be deceptive : Revealing a hidden viral infection with deep sequencing in plant quarantine context (Candresse *et al.*, 2014) (Annexe 2). Brièvement, les acides nucléiques viraux ont été amplifiés individuellement à l'aide d'étiquettes moléculaires (MID) sur des plaques de 96 puits. Les produits d'amplification d'une plaque sont tous groupés dans un seul tube. Au total, 32 tubes (soit 32 X 96 échantillons) ont été envoyés au pyroséquençage. Plusieurs raisons nous ont fait choisir le pyroséquençage (Roche 454) au début de ma thèse (2011) : la longueur des reads en comparaison des performances contemporaines de la technique Illumina, et la possibilité de multiplexer jusqu'à 96 échantillons par huitième de plaque 454 ce qui correspond à 768 échantillons par plaque. La combinaison de ces deux avantages nous semblait primordiale pour augmenter les

chances d'identifier les virus grâce aux algorithmes Blast, pour faire baisser le prix global de séquençage et pour analyser plusieurs milliers d'échantillons végétaux.

2.6. Traitement bioinformatique des séquences

Les données brutes issues du pyroséquençage (reads) ont tout d'abord été filtrées avec l'outil FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) qui contrôle la qualité des séquences et définit la longueur des séquences à analyser. Ces séquences filtrées ont ensuite été soumises au pipeline QIIME (Caporaso *et al.*, 2010) pour l'analyse des séquences. Les séquences ont ainsi été assignées à leur échantillon de provenance en fonction de leur MID. Ensuite, les MIDs, adaptateurs et amorces ont été enlevés de ces séquences. L'outil UCLUST (Edgar, 2010) implémenté dans QIIME a permis de regrouper les séquences en OTU (Unités Taxonomiques Opérationnelles) selon un seuil de similarité ≥97%. Pour chaque groupe de séquences un représentant a été choisi. Nous avons construit deux nouvelles bases de données virales (nucléotidique et protéique) par extraction de séquences à partir de la base NR (Non-redundant) du NCBI. Ainsi, un BlastN complété par un BlastX a été réalisé a partir de chaque représentant de groupe de séquences et le résultat du Blast (eValue : 0.001 et identité : 50%) a été attribué à chaque séquence du groupe du représentant. Toutes les données ont été regroupées dans une matrice de type table « biom ». Ces matrices contenant des séquences de virus animaux, végétaux et procaryotes, ont été triées pour ne garder que celles de virus végétaux et fongiques. Nous avons souhaité garder les virus de champignons car certaines familles contiennent des membres infectant des végétaux. Par ailleurs, sachant que divers virus sont étudiés en routine à BGPI (*Tomato leaf curl virus*, *Cauliflower mosaic virus*, *Sugarcane yellow leaf curl virus*, etc.), toutes les séquences présentant des hits avec des taux de similarité ≥97% pour ces virus issus du laboratoire ont été considérées comme des contaminations issues du laboratoire et ont été retirés de nos jeux de données.

2.7. Attribution des séquences aux genres et familles de virus de plantes

Pour chaque échantillon végétal, nous avons considéré que tous les hits viraux identifiés comme étant proches d'une même famille phytovirale (ou genre non assigné à une famille phytovirale) par les algorithmes BlastN/X étaient la trace de la présence d'un seul virus appartenant à la famille identifiée ou bien à un groupe taxonomiquement proche de cette famille. Cette approche, extrêmement conservative, réduit à un seul virus la présence potentielle de plusieurs virus appartenant à la même famille ou aux groupes taxonomiques proches. Par exemple, pour un échantillon végétal ayant 1000 reads dont les résultats du Blast correspondent à des *Geminiviridae* et 2 reads dont les résultats du Blast correspondent à des *Potyviridae*, nous avons considéré avoir détecté au moins un « *Geminiviridae*-like » et un « *Potyviridae*-like ». Nous avons arrêté notre analyse au niveau taxonomique de la famille car, comme cela a été évoqué dans la synthèse bibliographique de cette thèse, la recombinaison inter-genre n'est pas

anecdotique chez les virus de plantes, ce qui rend difficile l'attribution d'un genre viral à un read. Mener des analyses jusqu'au niveau de l'espèce serait encore plus périlleux car la recombinaison interspécifique est extrêmement fréquente chez plusieurs groupe de virus.

2.8. Calcul des indices de diversité α et β et courbes de raréfaction.

Les calculs d'index de diversité de Shannon-Wiener ont été effectués à l'aide du langage de programmation R via le package Vegan (Dixon, 2003). La diversité α a été estimée au niveau de la famille puis du genre pour la totalité des échantillons végétaux et au niveau de la famille pour les virus. La diversité α à l'échelle spécifique n'a pas été estimée car une grande partie des plantes échantillonnées n'ont pu être identifiées au niveau spécifique. Les échantillons végétaux et leurs virus associés ont ensuite été classés suivant les cinq types d'habitat (Type 0 à 4). Pour chaque habitat nous avons donc calculé les diversités α virales et végétales associées. Afin de comparer les diversités végétales et virales entre les différents habitats (diversité β) nous avons calculé l'indice de Morisita-Horn via le logiciel SPADE avec 200 bootstraps (Chao and Shen, 2010). Lorsque l'indice est inférieur à la valeur 1 les communautés sont différentes, cependant plus l'indice se rapproche de 1 plus les communautés sont similaires. Les courbes de raréfactions concernant les échantillonnages de végétaux ont quant à elles été tracées avec le logiciel Past3 (Hammer, 2001).

2.9. Calcul et comparaison des prévalences de communauté et nombre moyen de virus par échantillon virosé en fonction de divers paramètres écologiques

Afin de simplifier notre propos, nous avons choisi d'utiliser le terme « virosé » afin de désigner les échantillons végétaux sur lesquels au moins un read viral a été détecté. Ici, nous avons calculé les prévalences de communauté pour traiter du pourcentage d'échantillons virosés. Nous définissons la prévalence de communauté comme la proportion d'échantillons de la communauté végétale présentant des virus. Pour des raisons pratiques, nous utiliserons le terme « prévalence » dans la suite du manuscrit pour traiter de la prévalence de communauté. Le calcul de cette prévalence est donc le suivant :

$$Prévalence = \frac{Nombre\ d'échantillons\ végétaux\ virosés}{Nombre\ total\ d'échantillons végétaux} \times 100$$

Ainsi, nous avons calculé la prévalence globale pour chaque période d'échantillonnage mais aussi la prévalence pour chaque point de la grille d'échantillonnage. Des interpolations spatiales de la prévalence ont été réalisées pour chacun des quatre échantillonnages via le logiciel Surfer v9.0 en utilisant les paramètres par défaut (Anonymous, 1997).

Par ailleurs, nous avons calculé le nombre moyen de « virus » par échantillon virosé comme suit :

$$Nombre\ moyen\ de\ \text{virus}\ par\ échantillon\ virosé = \frac{Nombre\ total\ de\ virus\ détectés}{Nombre\ d'échantillons\ virosés}$$

Les prévalences et nombres moyens de « virus » par échantillon virosé ont été calculés pour chaque période d'échantillonnage de manière globale, puis pour les milieux cultivé et non-cultivé. Par ailleurs, la prévalence a également été calculée pour les paramètres écologiques suivants : plantes exotiques vs. plantes indigènes (uniquement pour le fynbos), plantes cultivées vs. plantes non-cultivées, et plantes pérennes vs. plantes annuelles, mais également pour les adventices vs. cultures au niveau des prairies, cultures diversifiées et habitats artificialisés (habitats de typologie 3 et 4).

Afin de comparer les prévalences deux à deux nous avons utilisé un test z permettant la comparaison de deux proportions. En ce qui concerne les nombres moyens de « virus » par échantillon virosé, nous avons tout d'abord testé la normalité des distributions du nombre de virus par échantillons virosés via un test de Shapiro-Wilk. Il est à noter qu'aucune de ces distributions n'étaient normales, nous avons alors effectué un test de Mann-Whitney afin de comparer les nombres moyen de « virus » par échantillon virosé. Ces tests statistiques ont été effectués avec le logiciel XLSTAT. Lorsque la p-value était ≥0.05 nous avons accepté l'hypothèse nulle (H0) propre à chaque test.

2.10. Analyse de la répartition spatio-temporelle des phytovirus

La distribution spatiale des phytovirus a été analysée via la méthode SADIE (Spatial Analysis by Distance IndicEs) (Perry, 1995; Perry, 1998). Cette méthode prend en compte des données de comptage sur chaque point de la zone d'échantillonnage. SADIE est implémentée dans un logiciel du même nom (Perry, 1996) qui calcule un indice d'agrégation *Ia* permettant d'évaluer le niveau global d'agrégation des individus (ici des virus). Lorsque $Ia>1$ il est habituellement considéré que la population est agrégée dans l'espace, si $Ia=1$ la population est répartie de manière aléatoire dans l'espace, et si $Ia<1$ alors la population est répartie de manière régulière dans l'espace. L'indice *Ia* peut être décomposé en deux index : un index d'agrégation positive v_i (indiquant le regroupement géographique des individus, e.g. « patchs ») et un index d'agrégation négative v_j (indiquant l'absence d'individus dans une zone géographique, e.g. « gaps »). Les données v_j et v_i sont plus informatives que l'indice brut *Ia*. La méthode se base sur l'hypothèse nulle selon laquelle les comptages observés sur l'aire d'échantillonnage sont répartis de manière aléatoire. Les indices d'agrégation *v* des virus ont été matérialisés sur une carte via le logiciel Surfer v9.0 (Anonymous, 1997) de

manière à visualiser les points où il y a une agrégation positive et ceux où elle est négative. Les contours ont été fixés à v_i=1,5 pour les patchs d'infection et v_j=-1,5 pour les gaps d'infection (Coutts *et al.*, 2004; Dader *et al.*, 2012; Jones, 2005; Jones *et al.*, 2008; Moreno *et al.*, 2007; Perry and Dixon, 2002; Perry *et al.*, 1999; Thackray *et al.*, 2002; Winder *et al.*, 2001), seuils arbitraires habituellement utilisés correspondant à 50% en plus que ce qui est attendu par chance (Perry *et al.*, 1999). Afin d'évaluer si les patrons détectés sont conservés de 2010 à 2012, nous avons calculé des indices d'association spatiale *X* basés sur les indices d'agrégation des deux années. Lorsque $X>0$ le patron est globalement conservé sur les deux années, lorsque $X<0$ le patron est globalement inversé d'une année à l'autre, et quand $X=0$ il n'y a aucune association spatio-temporelle. Lorsque la p-value associée aux indices d'agrégation est inférieure à 0,05 on considèrera que l'agrégation est significative, lorsque la p-value associée à *X* est <0,025 on considèrera que l'association est significative et lorsque la p-value est >0,975 on considèrera qu'il y a une dissociation significative. Ces indices d'association spatiale pour chacun des points de la grille d'échantillonnage ont également été matérialisés via le logiciel Surfer v9.0 (Anonymous, 1997).

2.11. Analyses multivariées

Nous avons réalisé des analyses factorielles discriminantes (AFD) s'appliquant à des données quantitatives (nombre de familles virales au niveau de chacun des 100 points d'échantillonnage) sur lesquelles est déjà définie une typologie ou partition, dans notre cas les 5 types d'habitats (variable qualitative). Cette analyse a pour objectif de représenter si les familles virales sont structurées en fonction du type d'habitat (habitats naturels, habitats naturels perturbés, habitats semi-naturels, cultures diversifiées et prairies et habitats artificialisés). Ces analyses ont été effectuées avec le logiciel XLSTAT en utilisant les paramètres par défaut.

2.12. Obtention de génomes entiers de *Geminiviridae*-like et réalisation d'une phylogénie

Une sélection de neuf échantillons à partir desquels des reads de *Geminiviridae*-like ont été obtenus pour F2010 a été soumise à une Rolling Circle Amplification (RCA) en utilisant le protocole décrit par Shepherd *et al.* (Shepherd *et al.,* 2008). Les produits de RCA ont été digérés soit par *Bam*HI ou *Sma*I ou *Eco*RI ou *Xmn*I pendant 3h à 37°C. Les fragments de 1.7 à 3kb ainsi générés ont été purifiés sur gel et clonés dans le vecteur pJET1.2 (Thermo Fisher, USA). Les clones ont été séquencés par la société Macrogen via la méthode Sanger. Les séquences de la protéine Rep (Replication–associated protein) de ces génomes ainsi obtenus et celle d'autres génomes de geminivirus ainsi que de génomes de virus proches des géminivirus (mycovirus, nanovirus, gemycircularvirus, etc.) ont été alignées via la méthode ClustalW (Thompson *et al.*, 1994) implémentée dans MEGA 5.2.1 (Tamura *et al.*, 2011). À partir de ces alignements un arbre phylogénétique à été réalisé selon la méthode de maximum de vraisemblance avec 500 bootstraps à l'aide du logiciel PhyML 3.1 (Guindon *et al.*, 2005) en ayant sélectionné au

préalable le meilleur modèle évolutif décrivant nos alignements via le logiciel JModelTest 2.1.6 (Posada, 2008). Nous avons également réalise des BlastN et BlastX afin de pouvoir déterminer l'identité taxonomique de ces génomes.

3. Résultats

3.1. Résultats issus du séquençage

Suite au séquençage nous avons obtenus 1 332 624 reads pour F2010, 1 259 767 reads pour F2012, 1 092 351 reads pour C2010 et 1 444 951 reads pour C2012 avec des longueurs moyennes de reads respectives de 247, 243, 300, et 259 nucléotides. En annexes 3 et 4 sont répertoriées les données brutes concernant les reads issus des NGS. On peut observer que les distributions de reads sont similaires pour les échantillonnages F2010, F2012 et C2012, en effet un grand nombre de reads se situe autour de la moyenne (entre 160 et 300pb) alors que pour l'échantillonnage C2010, on a une distribution plus large de la taille des reads avec un pic (entre 400 et 500pb) qui se situe au-delà de la moyenne qui est de 300pb.

3.2. Diversité globale des communautés végétales

Le nombre d'échantillons récoltés dans le fynbos est plus important que celui de Camargue (Tableau 1.1). Par ailleurs, l'échantillonnage C2010 contient 135 échantillons de moins que C2012. Quatre-vingt onze à 98% de nos échantillons ont pu être identifiés au niveau de la famille, 89 à 97% au niveau du genre et 55 à 86% au niveau de l'espèce (Tableau 1.1).

		Identification						Diversité (Indice de Shannon-Wiener)	
		Famille		Genre		Espèce			
	Nombre total d'échantillons	Nombre	Pourcentage	Nombre	Pourcentage	Nombre	Pourcentage	Famille	Genre
F2010	743	676	91	662	89	47	64	3,09	4,19
F2012	713	673	94	640	90	393	55	3	4,2
C2010	495	469	95	461	93	404	82	2,1	3,65
C2012	630	618	98	611	97	541	86	2,42	3,84

Tableau 1.1 : Tableau indiquant le nombre d'échantillons végétaux récoltés et la proportion de plantes identifiées au niveau de la famille, du genre, et de l'espèce ainsi que l'indice de diversité de Shannon-Wiener au niveau de la famille et du genre pour chaque période d'échantillonnage.

Afin de savoir si nous avions fourni un effort d'échantillonnage permettant de capturer toute la diversité de chaque écosystème nous avons tracé des courbes de raréfaction que ce soit au niveau de la famille ou du genre (Figure 1.5). Si l'on se situe au niveau du genre, quelque soit l'échantillonnage il semble que l'effort n'ait pas été suffisant, bien que la pente de la courbe diminue quand le nombre d'individus échantillonné devient important. Pour les échantillonnages concernant le fynbos il semblerait que nous ayons capturé l'essentiel de la diversité en terme de famille, par contre ce n'est pas totalement le cas pour l'échantillonnage de Camargue qui présente le même type de courbe que pour les genres.

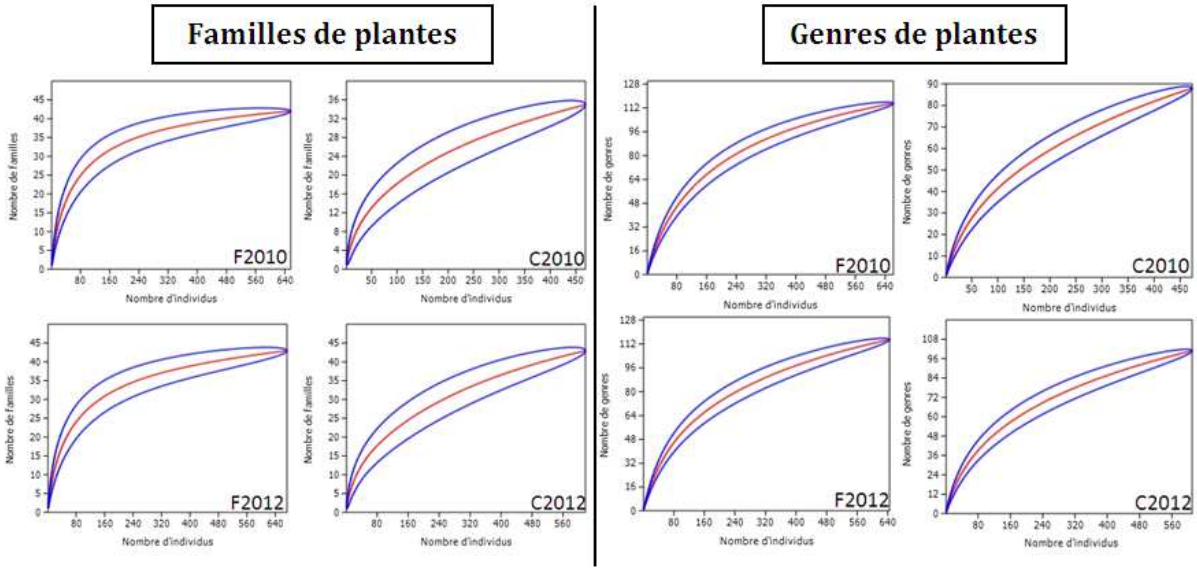


Figure 1.5 : Courbes de raréfaction des communautés (en rouge) de plantes échantillonnées selon la famille et le genre. L'intervalle de confiance à 95% est représenté en bleu.

Par ailleurs, le fait qu'une grande partie des espèces n'ait pu être identifiée peut engendrer une perte d'information concernant le calcul d'indices de diversité. Ainsi nous avons calculé l'indice de Shannon-Wiener au niveau de la famille et du genre pour chaque période d'échantillonnage (Tableau 1.1). Que ce soit au niveau de la famille ou du genre des plantes échantillonnées, l'indice de Shannon-Wiener est plus élevé pour les échantillonnages effectués dans le fynbos que pour la Camargue. Cela semble indiquer que les communautés de plantes échantillonnées dans le fynbos sont globalement plus diversifiées que les communautés échantillonnées en Camargue. Les familles ainsi que les genres de plantes de chaque campagne d'échantillonnage ont été comparés deux à deux via l'indice de Morisita-Horn (M) afin d'évaluer si les communautés étaient différentes entre les années d'échantillonnage et entre les lieux géographiques. La diversité botanique du fynbos est ainsi différente de la diversité botanique en Camargue (0.52<M<0,66, rang taxonomique de la famille), ce qui était fortement attendu. En revanche, les jeux de plantes échantillonnés en 2010 et 2012 sont très similaires dans chaque région que ce soit en Camargue ou en Afrique du Sud (0.93<M<0,98, rang taxonomique de la famille) (Tableau 1.2).

	Indice de Morisita-Horn	
Echantillons comparés	Familles de plantes	Genres de plantes
F2010 vs. F2012	0,9346 (±0,0196)	0,8039 (±0,0260)
F2010 vs. C2010	0,5489 (±0,0337)	0,0581 (±0,0142)
F2010 vs. C2012	0,5247 (±0,0371)	0,0654 (±0,0130)
F2012 vs. C2010	0,6648 (±0,0379)	0,0655 (±0,0142)
F2012 vs. C2012	0,6445 (±0,0377)	0,0759 (±0,0138)
C2010 vs. C2012	0,9818 (±0,0073)	0,8833 (±0,0248)

Tableau 1.2 : Comparaison de la composition des communautés de plantes au niveau du genre et de la famille via l'indice de Morisita-Horn (±SE).

Nous avons ensuite calculé les indices de diversité de Shannon-Wiener de chaque communauté végétale (au niveau du genre et de la famille) pour chaque habitat (score attribué à l'habitat variant de 0 à 4). En cohérence avec les effets attendus de l'anthropisation, on observe dans tous les cas une diminution graduelle de la diversité des communautés avec l'augmentation du niveau d'anthropisation du milieu (Figure 1.6).

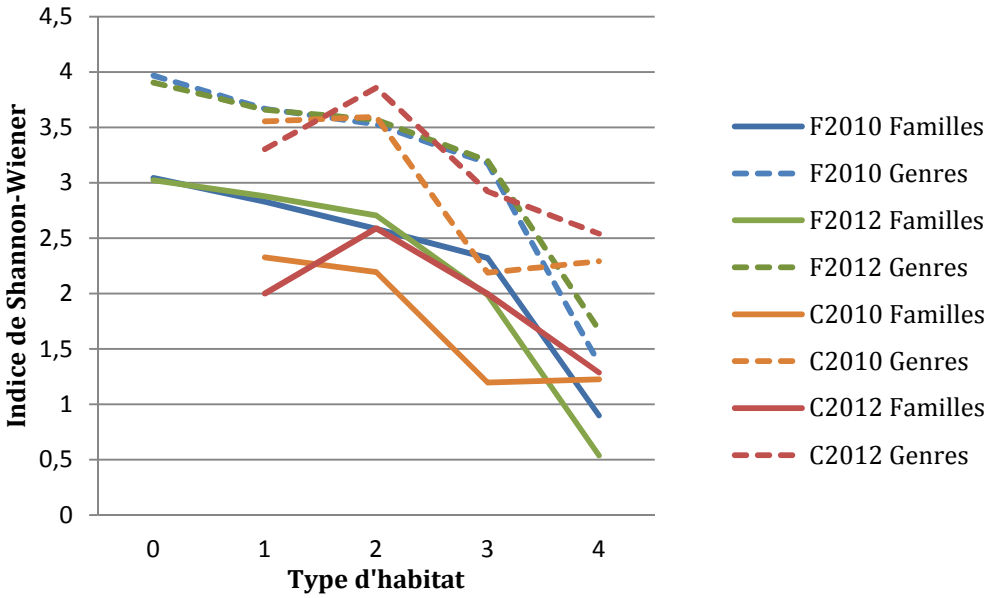


Figure 1. 6 : Graphique représentant les indices de Shannon-Wiener pour les genres et familles des végétaux des différents échantillonnages en fonction du type d'habitat dans lesquels ils ont étés récoltés. 0=Habitats naturels, 1=Habitats naturels perturbés, 2=Habitats semi-naturels, 3=Cultures diversifiées et prairies, 4=Habitats artificialisés.

3.3.Diversité virale, prévalence de communauté et nombre moyen de « virus » par échantillon virosé

Bien que le nombre de représentants de familles virales dans les 4 échantillonnages varie de 223 à 687, les indices de Shannon-Wiener associés à chaque communauté végétale sont très similaires (2.44 à 2.58 ; Tableau 1.3). Le pourcentage d'échantillons végétaux sur lesquels on a détecté au moins un read correspondant à un virus de plante ou de champignon varie de 25.95% (F2012) à 58,41% (C2012), les nombres moyens de « virus » par échantillon végétal virosé nous indiquent que les échantillons virosés ont en moyenne 1,6 entités virales (à l'échelle de la famille virale). Par ailleurs, nous avons calculé la co-occurrence virale au sein de nos échantillons qui est définie comme étant la présence d'au moins deux virus appartenant à des familles différentes dans un même échantillon. Cette co-occurrence virale varie de 8.27%, à 30.95% (Tableau 1.3).

Echantillonnage	Nombre de "virus"	Indice de Shannon-Wiener	Prévalence (%)	Nombre moyen de "virus" par échantillon virosé	Co-occurrence virale (%)
F2010	448	2,45	35,80	1,68	15,30
F2012	261	2,44	25,95	1,41	8,27
C2010	223	2,58	33,54	1,34	8,89
C2012	688	2,52	58,41	1,87	30,95

Tableau 1.3 : Tableau récapitulant les informations sur le nombre total de représentants de familles virales détectés par période d'échantillonnage (nombre de « virus ») et indices de Shannon-Wiener associés. La co-occurrence virale correspond au nombre d'échantillons végétaux sur lesquels ont été détectés au moins 2 virus sur le nombre total d'échantillons végétaux x100.

Globalement, 19 familles de virus de plantes sur les 23 décrites par l'ICTV ont été identifiées (Figure 1.7), dont 18 pour C2012, 17 pour F2010, 16 pour F2012, et 12 pour C2010. On remarquera que les virus à ARN sont beaucoup plus représentés que les virus à ADN avec des proportions respectives de 85%, 57%, 75% et 84% pour F2010, F2012, C2010 et C2012. De plus, ces virus à ARN sont majoritairement représentés par les virus à ssRNA+ et dsRNA. Suivant les échantillonnages, les proportions de chaque famille virale varient mais l'on constate tout de même une forte représentation des *Closteroviridae*-like quelque soit l'échantillonnage. Pour F2012 on observe une augmentation de la prévalence des *Caulimoviridae*-like. Nous ne fournirons pas de résultats au niveau du genre, car comme précisé précédemment, les évènements de recombinaison peuvent biaiser les résultats.

L'analyse de similarité des communautés virales de chaque échantillonnage nous a donné des indices de Morisita-Horn allant de 0.91 (SE=0.0186) pour la comparaison F2010 vs. C2012 à 0.68 pour la comparaison F2012 vs. C2010 (SE=0.0562) (Tableau 1.4). Ces résultats montrent que les familles phytovirales ne sont pas structurées géographiquement comme le sont les familles botaniques, e.g. la communauté virale de F2010 est plus proche de celle de C2012 que de celle de F2012 (M=0.7074, SE=0.0478, et M=0.9122, SE=0.0186 respectivement).

	Indice de Morisita-Horn
Echantillons comparés	Familles virales
F2010 vs. F2012	0,7074 (±0,0478)
F2010 vs. C2010	0,8691 (±0,0364)
F2010 vs. C2012	0,9122 (±0,0186)
F2012 vs. C2010	0,684 (±0,0562)
F2012 vs. C2012	0,7152 (±0,0478)
C2010 vs. C2012	0,8486 (±0,0362)

Tableau 1.4 : Comparaison de la composition des communautés virales au niveau de la famille via l'indice de Morisita-Horn (±SE).

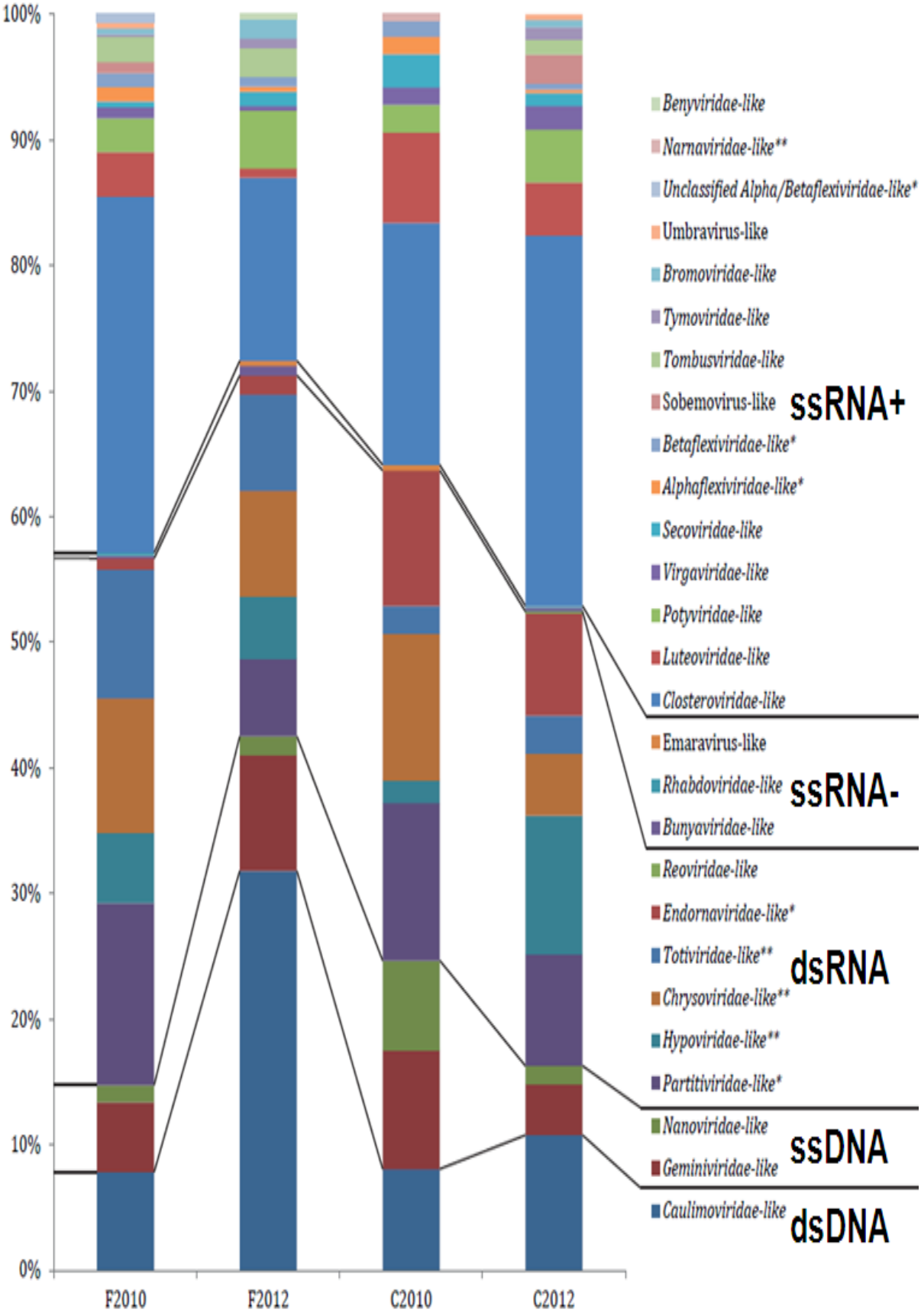


Figure 1.7 : Répartition des représentants de familles au sein des différentes familles de virus de plantes (et de genres non assignés à une famille) et de champignons. *famille contenant des virus de champignons, **famille contenant uniquement des virus de champignons.

3.4. Diversités virales, prévalences et nombres moyens de « virus » par échantillon virosé à l'échelle des milieux cultivé et non-cultivé

Le Tableau 1.5 récapitule les prévalences virales moyennes obtenues en fonction du milieu ainsi que les indices de diversité au niveau taxonomique de la famille associés aux communautés virales et végétales. À l'exception de l'échantillonnage C2012 pour lequel l'indice de Shannon-Wiener du milieu cultivé est légèrement au dessus de celui du milieu non-cultivé, les indices de Shannon-Wiener montrent globalement que les communautés végétales du milieu cultivé sont moins diversifiées que celles du milieu non-cultivé. L'indice de diversité de Morisita-Horn indique en outre que ces communautés sont différentes, bien que les communautés végétales du milieu cultivé soient plus proches de celles du milieu non-cultivé en Camargue en comparaison de celles du fynbos pour lequel les communautés sont fortement dissimilaires.

				Plantes (Familles)		Virus (Familles)	
		Pourcentage d'échantillons virosés	Nombre de "virus" par échantillon virosé	Indice de Shannon-Wiener	Indice de Morisita (±SE)	Indice de Shannon-Wiener	Indice de Morisita (±SE)
F2010	Cultivé	47,83*	1,89*	2,45	0,32 (±0,04)	2,406	0,88 (±0,05)
	Non-Cultivé	31,79*	1,58*	3,03		2,395	
F2012	Cultivé	31,55*	1,46*	2,25	0,44 (±0,05)	2,363	0,91 (±0,04)
	Non-Cultivé	23,95*	1,69*	3,04		2,431	
C2010	Cultivé	34,71	1,36	1,7	0,78 (±0,06)	2,217	0,74 (±0,08)
	Non-Cultivé	30,77	1,32	2,33		2,39	
C2012	Cultivé	61,32*	1,85	2,1	0,96 (±0,06)	2,395	0,97 (±0,02)
	Non-Cultivé	52,43*	1,92	2		2,389	

Tableau 1.5 : Tableau récapitulatif des prévalences virales, du nombre moyen de « virus » par échantillon virosé et des indices de diversité concernant les plantes et les virus en fonction du type de milieu. Concernant les prévalences, les différences significatives entre milieu cultivé et non cultivé sont indiqués par une étoile (test z, $p<0,05$). Les différences significatives entre les nombres moyens de « virus » par échantillon virosé sont également indiqués par une étoile (test de Mann-Withney, $p<0.05$).

Les prévalences virales des milieux cultivés sont supérieures à celles du milieu non-cultivé pour tous les échantillonnages; la différence est significative pour les échantillonnages F2010, F2012 et C2012 ($p<0.05$, Tableau 1.5). Par ailleurs, si l'on sépare nos échantillons en fonction du fait que les plantes sont cultivées (riz, blé, luzerne, etc.) ou non (adventices des cultures et plantes sauvages), on observe que les prévalences sont bien supérieures dans les cultures pour tous les échantillonnages, et ceci de façon significative pour F2010, F2012 et C2012 ($p<0.05$, Figure 1.8). Concernant le nombre moyen de « virus » par échantillon virosé (Tableau 1.5), des différences significatives ($p<0.05$) ont été détectées pour les échantillonnages réalisés dans le fynbos, toutefois les tendances sont inversées : en 2010 le nombre moyen de « virus » par échantillon virosé est significativement plus élevé en milieu cultivé, alors qu'en 2012 il est plus élevé en milieu non-cultivé. Concernant la diversité des virus associés à

chaque milieu (indices de Shannon-Wiener, Tableau 1.5), on remarque que les indices de Shannon-Wiener sont très proches et l'indice de Morisita-Horn indique que les communautés virales des milieux cultivé et non-cultivé sont globalement similaires (M>0,7). Ce résultat qui confirme les résultats obtenus dans la partie précédente semble indiquer que les familles virales ne sont globalement pas inféodées à un milieu particulier et à sa végétation associée.

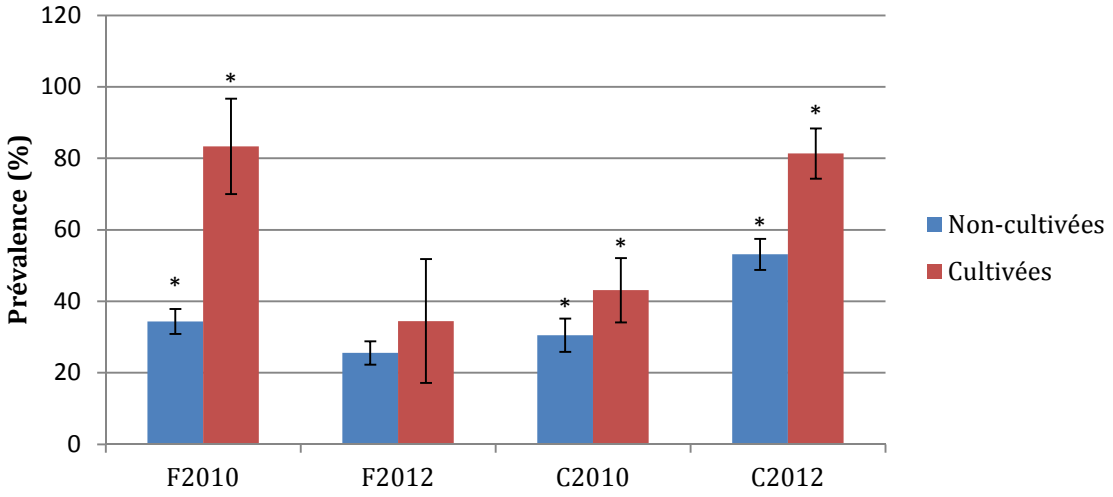


Figure 1.8 : Proportions d'échantillons virosés cultivés vs. non-cultivés pour les différentes périodes d'échantillonnage. Les étoiles indiquent une différence significative (z-test, $p<0.05$ entre les proportions comparées deux à deux).

Les interpolations spatiales du pourcentage d'échantillons positifs à chaque période d'échantillonnage (Figure 1.9) permettent d'illustrer les résultats analytiques précédents. On observe en effet une prévalence virale plus élevée dans le milieu cultivé pour les collectes F2010 et C2012. Cependant, le patron de prévalence plus élevée en milieu cultivé pour F2012 est plus difficile à déceler visuellement, cela peut être expliqué par des prévalences plus faibles avec une distribution globale des virus non agrégée. On remarque par ailleurs que le jeu de données C2010 ne semble pas être caractérisés par une dichotomie zones cultivées / zones non-cultivées. Des « hotspots » de prévalence virale sont par exemple détectés dans les deux types de milieux pour C2010.

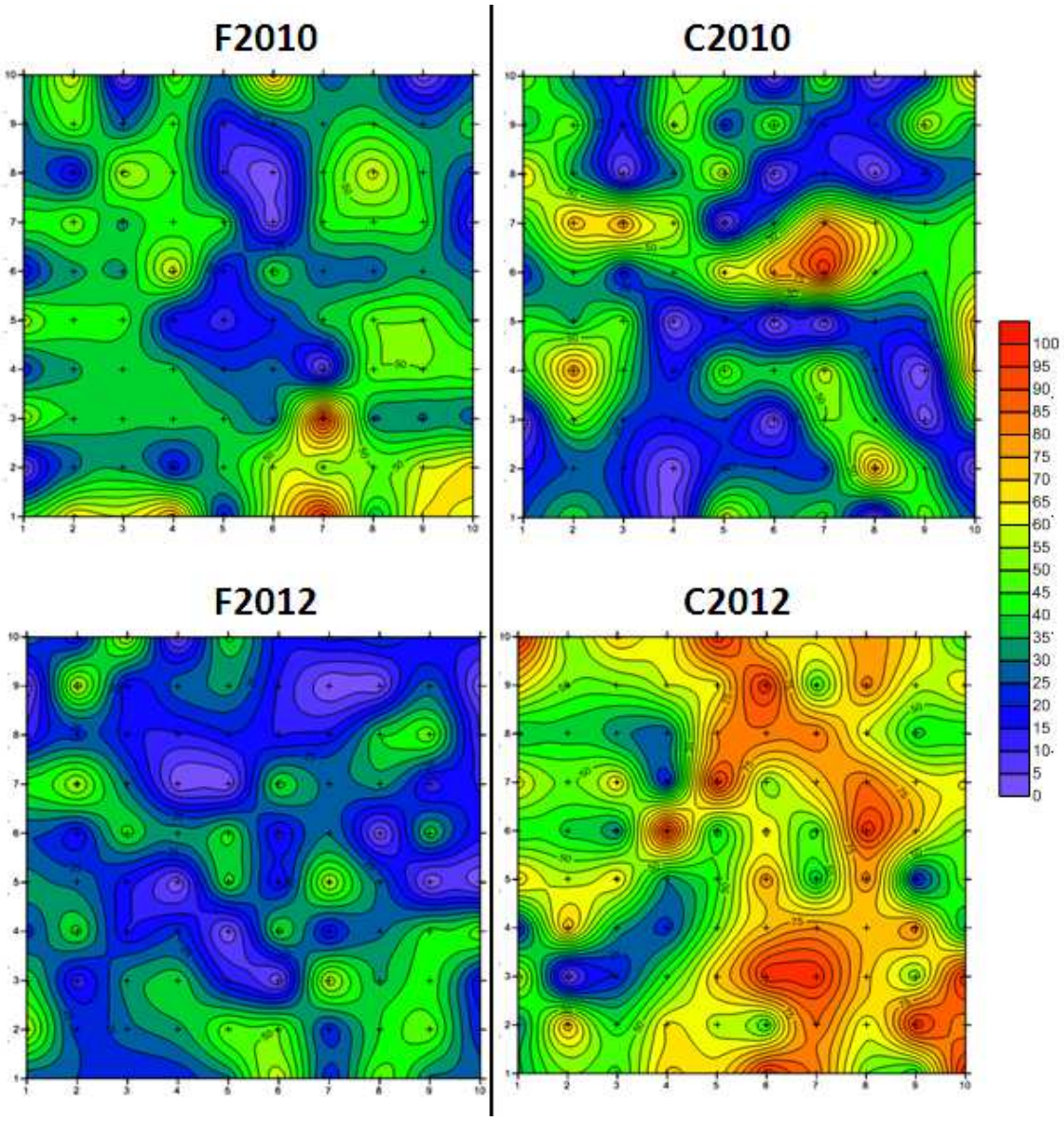


Figure 1.9 : Interpolations spatiales du pourcentage d'échantillons pour lesquels la présence de virus de plantes et de champignons a été détectée. L'échelle indiquant le pourcentage d'échantillonssur lesquelles des virus de plantes et de champignons ont été détectés est représentée à droite des grilles.

3.5. Comparaison des prévalences virales et du nombre moyen de « virus » par échantillon virosé en fonction du « statut » des plantes hôtes (pérennes vs. annuelles, adventices vs. cultivées, exotiques vs. indigènes)

Au-delà de l'influence globale du type de milieu sur les prévalences virales, nous avons souhaité analyser les prévalences virales en fonction du statut des plantes hôtes. Pour cela nous disposions de données concernant leur mode de vie (plantes pérennes ou annuelles, les plantes ayant des modes de vie mixtes n'ont pas été traitées), leur origine géographique (plantes indigènes ou exotiques) et leur identité dans le milieu cultivé (plantes cultivées ou adventices).

3.5.1. Plantes pérennes vs. plantes annuelles

Les proportions d'échantillons de plantes pérennes et annuelles virosés ne sont pas significativement différentes pour les échantillonnages F2012, C2010, C2012 ($p>0.05$, Figure 1.10a). En revanche, pour F2010, la proportion d'échantillons de plantes annuelles virosés est significativement plus importante que celle des échantillons de plantes pérennes ($p<0.0001$) (Figure 1.10a). Afin de savoir si la tendance était la même dans les milieux cultivé et non-cultivé, nous avons comparé la proportion d'échantillons de plantes annuelles et de plantes pérennes virosés dans chacun des deux compartiments. Pour les échantillonnages F2012, C2010, et C2012, les proportions d'échantillons de plantes pérennes et annuelles virosés ne sont pas significativement différentes ($p>0.05$) alors que pour F2010, la proportion d'échantillons de plantes annuelles cultivées virosés est toujours plus importante que la proportion d'échantillons de plantes pérennes cultivées virosés ($p<0.05$) que l'on se situe dans le milieu cultivé ou non-cultivé.

Les nombres moyens de « virus » par échantillon végétal virosé ne sont pas significativement différents entre les échantillons de plantes pérennes et de plantes annuelles dans les échantillonnages F2010, F2012 et C2012 ($p>0.05$). Toutefois, le nombre moyen de « virus » par échantillon virosé est significativement plus élevé pour les annuelles de C2010 ($p<0.05$) (Tableau 1.6).

3.5.2. Adventices vs. plantes cultivées

Si l'on se focalise maintenant sur les proportions des échantillons de plantes adventices/cultivées virosés dans les cultures diversifiées, prairies, et habitats artificialisés, les prévalences virales des échantillons de plantes cultivées sont significativement plus élevées pour F2010 et C2012. De façon intéressante, la tendance est inversée pour C2010 (Figure 1.10b). Les nombres moyens de « virus » par échantillons de plantes virosés ne sont quant à eux pas significativement différents entre plantes cultivées et adventices.

3.5.3. Plantes exotiques vs. plantes indigènes

A l'échelle globale de l'agro-ecosystème de Buffelsfontein et pour les deux années de collecte, les prévalences virales des échantillons de plantes exotiques sont significativement plus élevées que celles des échantillons de plantes indigènes ($p<0.05$, Figure 1.10c). Les nombres moyens de « virus » par échantillon de plantes virosé ne sont en revanche pas significativement différents (Tableau 1.6). Il est par ailleurs très intéressant de noter que les proportions d'échantillons de plantes virosés exotiques et indigènes ne sont pas significativement différentes dans le milieu cultivé (F2010 et F2012) alors que pour une année de collecte sur les deux (F2010), la prévalence des échantillons de plantes exotiques virosés situés dans les zones non-cultivées était significativement plus importante que la prévalence des échantillons de plantes indigènes virosés ($p<0.05$).

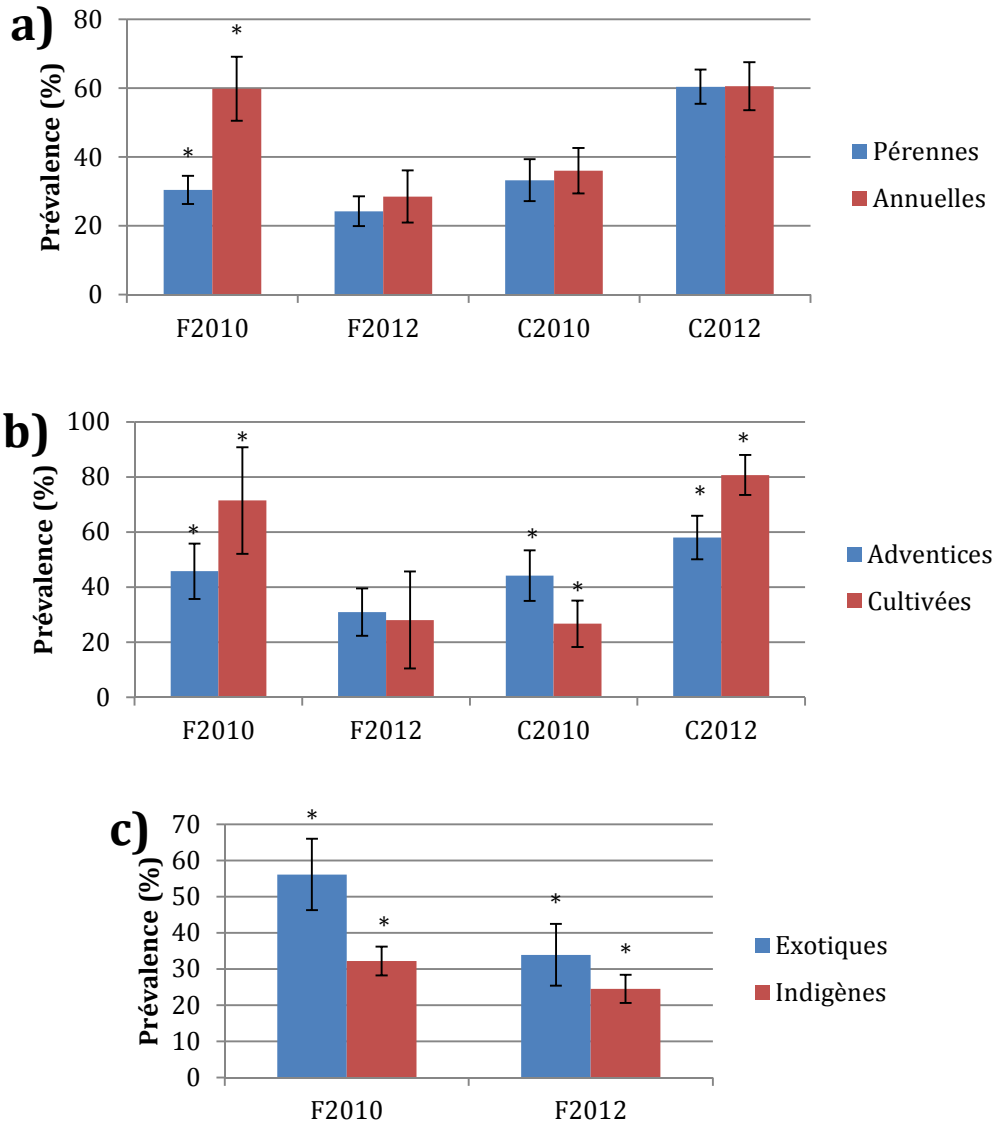


Figure 1.10 : Proportion d'échantillons de plantes virosés en fonction de différents paramètres concernant les plantes des différentes périodes d'échantillonnage. a) Proportions d'échantillons de plantes pérennes vs. annuelles virosés b) Proportions d'échantillons de plantes adventices vs. cultivées virosés dans les habitats artificialisés, cultures diversifiées et prairies (habitats 3 et 4) c) Proportions d'échantillons de plantes exotiques vs. indigènes virosés. Les étoiles indiquent une différence significative (z-test, $p<0.05$ entre les proportions comparées deux à deux).

Echantillonnage	Pérennes	Annuelles	Cultivées	Non-cultivées	Exotiques	Indigènes
F2010	1,73	1,50	1,67	1,81	1,84	1,60
F2012	1,41	1,49	1,33	1,30	1,34	1,46
C2010	1,47*	1,19*	1,32	1,23		
C2012	2,00	1,71	1,88	1,96		

Tableau 1.6 : Tableau récapitulant les nombres moyens de « virus » par échantillon de plantes virosée pour les plantes pérennes, annuelles, cultivées, non-cultivées, exotiques et indigènes. Les étoiles indiquent une différence significative (Test de Man-Whitney, $p<0.05$) entre deux niveaux d'infections comparés deux à deux.

3.5.4. Répartition spatio-temporelle des phytovirus

Les analyses spatio-temporelles réalisées à l'aide de l'algorithme SADIE montrent que la majorité des familles virales sont réparties de manière aléatoire en Afrique du Sud et en France en 2010 et en 2012, c'est-à-dire qu'elles ne sont pas spatialement agrégées. Ces résultats ne sont cependant pas significatifs pour la majorité des familles, mettant probablement en évidence la faiblesse des effectifs. Toutefois, quelques familles virales présentent des patrons significatifs d'agrégation. Par exemple, les *Luteoviridae*-like semblent inféodés au milieu cultivé pour les échantillonnages F2010 et C2012 (Figure 1.11). Si on se réfère aux plantes sur lesquelles ces virus ont été détectés, que se soit pour F2010 ou C2012, la majorité des échantillons de plantes virosés appartient à la famille des *Poaceae* et des *Fabaceae*, cependant seulement 5/16 (C2010) et 6/29 de ces échantillons plantes virosés sont cultivées (blé, avoine, trèfle). On peut donc supposer ici un rôle important des plantes sauvages et des adventices en tant qu'hôte de ces *Luteoviridae*-like, et donc éventuellement de réservoir de ces virus. Les *Geminiviridae*-like semblent quant à eux inféodés au milieu cultivé pour F2010 alors qu'on les retrouve dans le milieu non-cultivé pour C2010 (Figure 1.11). Les échantillons de plantes sur lesquelles ont été détectés ces virus appartiennent à diverses familles dont la majorité (23/25 pour F2010 et 19/21 pour C2010) sont des plantes adventices ou sauvages (brome, renoncule, etc.). Comme pour les *Luteoviridae*-like, les plantes adventices mais aussi sauvages pourraient avoir un rôle de réservoir pour les *Geminiviridae*-like. Ces deux familles virales apparaissent de manière épisodique, c'est-à-dire que le patron spatial n'est pas conservé sur les deux années. Ces épisodes épidémiques suggèrent que ces virus sont non-persistants à l'échelle de l'agro-écosystème et sont probablement transmis par des vecteurs.

Parmi les familles qui ont des patrons conservés sur les deux années d'échantillonnage on retrouve les *Partitiviridae*-like (dans le fynbos et en Camargue), les *Closteroviridae*-like dans le fynbos et les *Endornaviridae*-like en Camargue. Cependant, les *Partitiviridae*-like ont diminué en effectif de 2010 (N=65) à 2012 (N=16) dans le fynbos. De plus, ils étaient répartis dans les deux types de milieux en 2010 alors qu'en 2012 ils sont répartis dans le milieu non-cultivé (Figure 1.12). Toutefois on observe une conservation globale du patron sur 2 ans (X=0.2175, p=0.0269). En effet, si l'on se fie à la carte des indices d'association pour le fynbos, la zone centrale reste indemne de *Partitiviridae*-like et on observe toujours des zones d'infection à l'Ouest (milieu non-cultivé) et au Sud-Ouest bien qu'elles aient diminuées. Pour les échantillonnages en Camargue, on observe un cantonnement au milieu non-cultivé en 2010, alors qu'en 2012 ces virus semblent s'être propagés au milieu cultivé (en 2012 les indices d'agrégation indiquent qu'il n'y a plus d'agrégation spatiale) (Figure 1.12). Ainsi, en Camargue, les effectifs de *Partitiviridae*-like sont passés de 28 (C2010) à 61 (C2012). On observe ainsi une conservation du patron spatial à l'Ouest (milieu non-cultivé), ainsi qu'au Sud-Est qui reste une zone sans présence virale. Concernant les plantes sur lesquelles ont été détectés ces *Partitiviridae*-like, ce sont majoritairement des plantes non-cultivées, il ne semble pas y avoir de tendance à des infections préférentielles sur les plantes pérennes,

en effet on observe respectivement 40%, 20%, 64% et 79% d'échantillons de plantes pérennes virosées pour les échantillonnages F2010, F2012, C2010 et C2012.

La deuxième famille de virus de plantes qui a des patrons conservés sur les deux années d'échantillonnage est les *Endornaviridae*-like (Figure 1.13). Contrairement à la Camargue où la prévalence de ces virus était suffisante pour réaliser des analyses spatiales, dans le fynbos seuls 4 représentants ont été identifiés sur les deux échantillonnages. Il est intéressant de noter que, bien que cette famille soit largement prévalente sur les deux années de collecte en Camargue, l'indice d'association indique que les patrons ne sont globalement pas conservés (Figure 1.13). Si l'on s'intéresse plus précisément aux échantillons de plantes à partir desquels ont été détectés des *Endornaviridae*-like, on se rend compte que 75% (2010) et 81.5% (2012) des échantillons concernées sont du riz.

La dernière famille pour laquelle les effectifs sont importants est les *Closteroviridae*-like. Dans le fynbos, les *Closteroviridae*-like ne sont pas agrégés significativement en 2010, bien qu'on les retrouve majoritairement dans le milieu non-cultivé. En 2012, leur effectif diminue (de 127 en 2010 à 38 en 2012) et ils sont alors agrégés dans le milieu non-cultivé (Figure 1.14). L'association spatiale les concernant est très significative sur les deux années, en effet, les zones d'infection dans le milieu non-cultivé de l'Ouest du dispositif sont très conservées, alors que les zones du Sud-Est qui correspondent aux cultures intensives restent indemnes (Figure 1.14). Au total, les *Closteroviridae*-like ont été détectés sur 25 familles végétales indiquant ainsi leur large gamme d'hôte. De plus 69% (F2010) et 73% (F2012) des échantillons de plantes virosées sont des plantes pérennes.

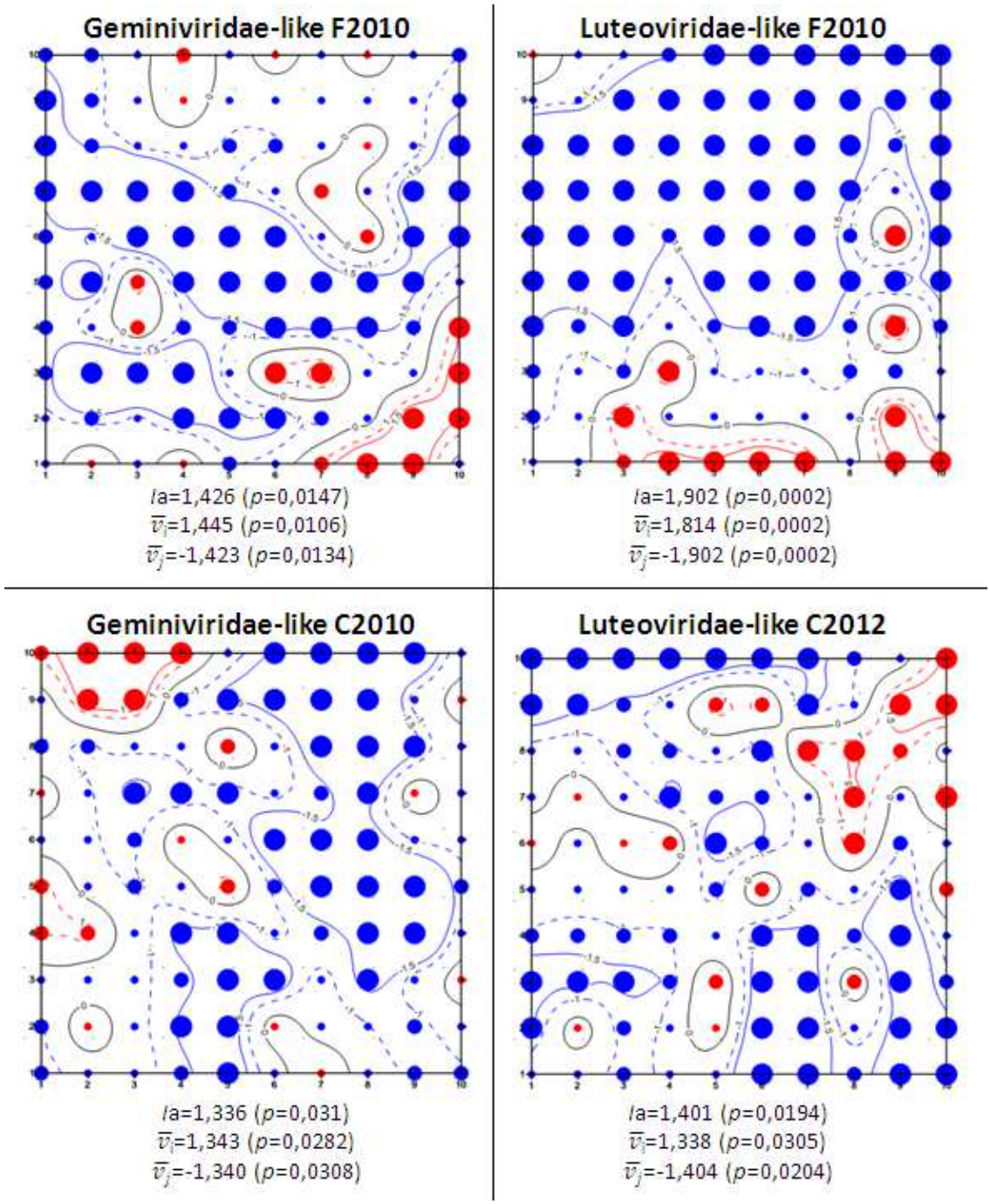


Figure 1.11 : Cartes représentant les indices d'agrégation concernant les *Geminiviridae*-like (F2010 et C2010) et les *Luteoviridae*-like (F2010 et C2012) sur les grilles d'échantillonnage (SADIE). Les points représentent chaque point d'échantillonnage, lorsque $v>0$ on est en présence d'un « patch » d'infection (rouge), lorsque $v<0$ on est en présence d'un « gap » d'infection (bleu). Les petits points représentent des indices d'agrégation allant de 0 à ±0,99 (agrégation plus faible que celle attendue), les points moyens représentent les indices d'agrégation allant de ±1 à ±1,49 (agrégation légèrement au dessus de l'attendu), et les grands points représentent les indices d'agrégation >1,5 ou <1,5 (agrégation qui est une moitié de plus que celle attendue par chance). Les lignes rouges en pointillé délimitent v=1, les lignes rouges pleines v=1.5, les lignes bleues en pointillé v=-1, et les lignes bleues pleines v=-1,5. La ligne noire correspond à v=0, elle délimité les « patchs » et les « gaps » d'infection. L'indice d'agrégation Ia et les indices moyens d'agrégation en « patchs » et en « gaps » ($\overline{v}_i$ et $\overline{v}_j$) sont indiqués en dessous des cartes accompagnés de leur p-value.

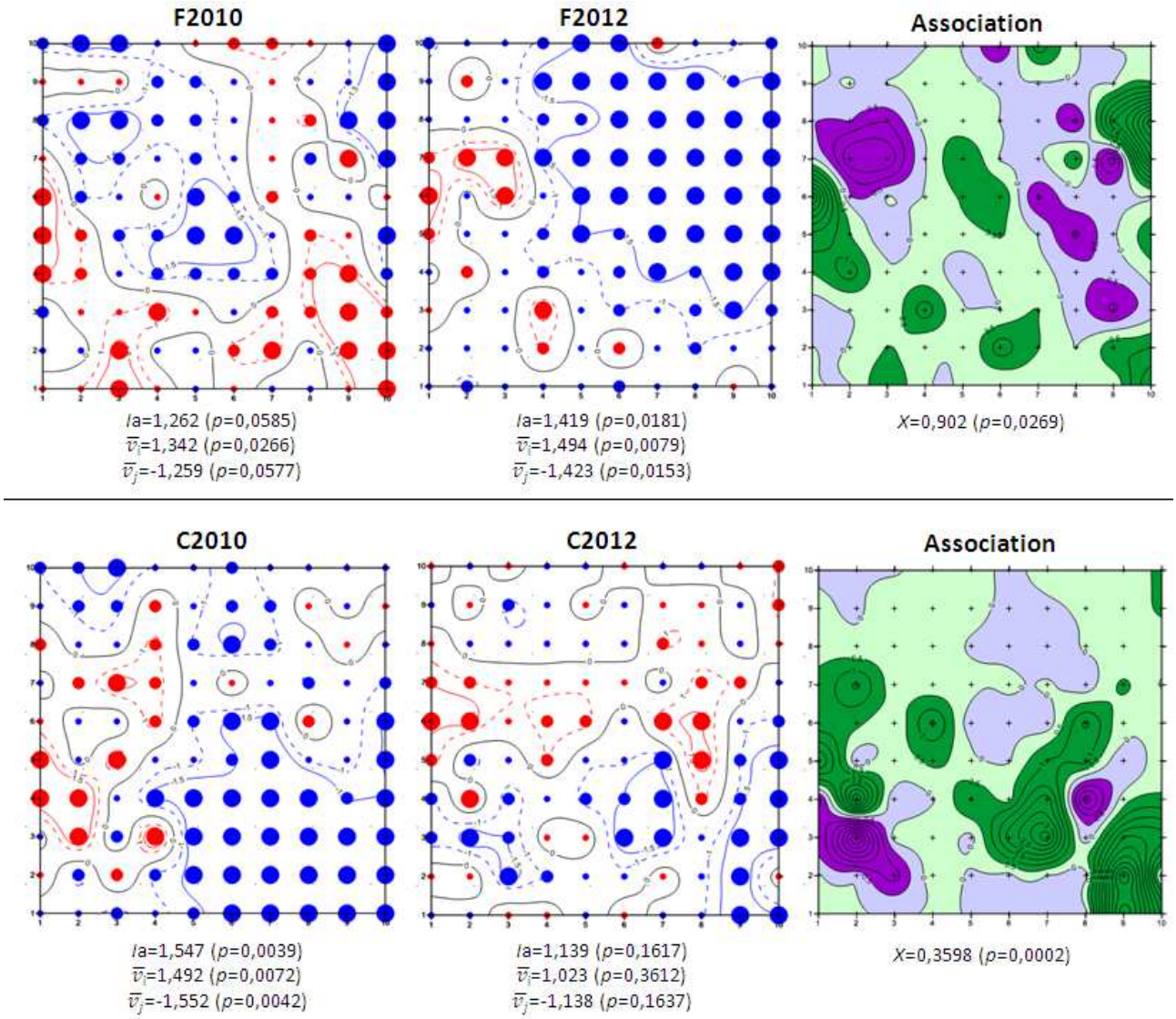


Figure 1.12 : Cartes représentant les indices d'agrégation et l'association spatiale concernant les *Partitiviridae*-like sur les grilles de chaque période d'échantillonnage. Se référer à la Figure 1.11 pour la légende. Les indices d'association spatiale sur les deux années sont représentés sur les cartes nommées « Association ». Les niveaux d'association spatiale sont symbolisés par les lignes noires, la couleur vert foncé indique que X>0,5 (le patron est globalement conservé sur les deux années) la couleur violette indique que X<-0,5 (le patron est globalement inversé), la couleur mauve indique que -0,5<X<0 et la couleur vert clair indique que 0<X<0,5.

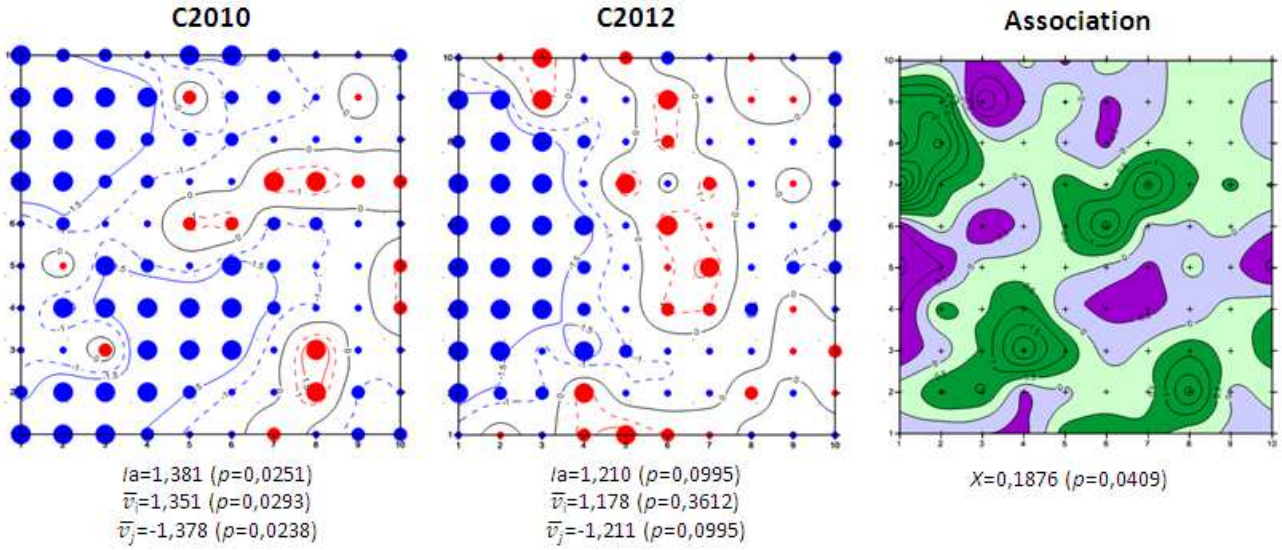


Figure 1.13 : Cartes représentant les indices d'agrégation et l'association spatiale concernant les *Endornaviridae*-like sur la grille d'échantillonnage en Camargue. Se référer aux Figures 1.11 et 1.12 pour la légende.

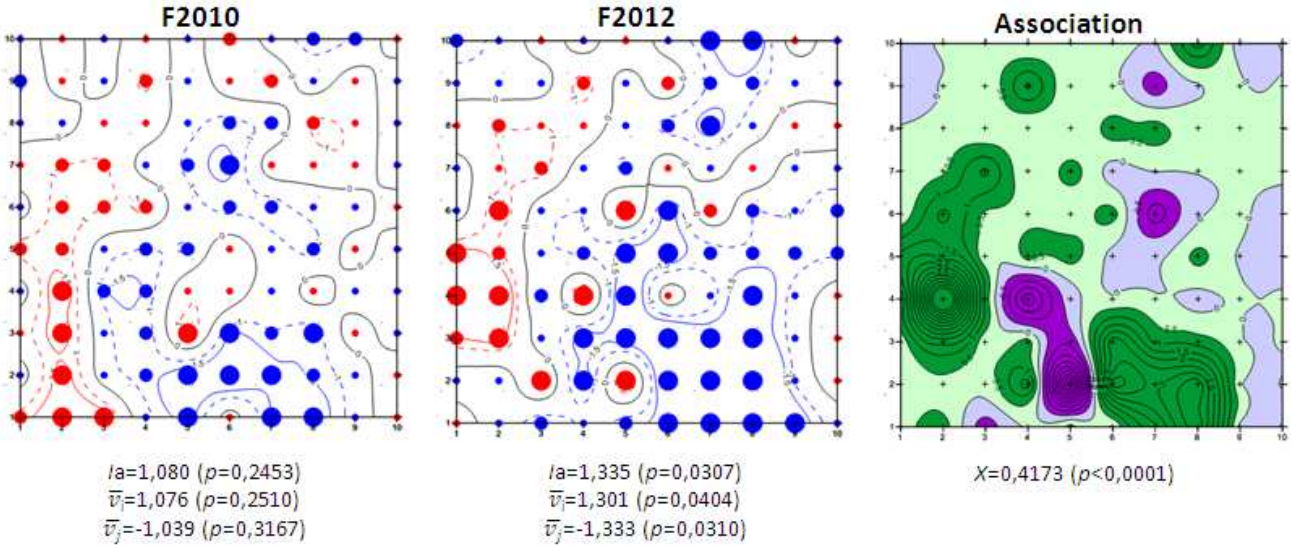


Figure 1.14 : Cartes représentant les indices d'agrégation et l'association spatiale concernant les *Closteroviridae*-like sur la grille d'échantillonnage du fynbos. Se référer aux Figures 1.11 et 1.12 pour la légende.

3.5.5. Analyses factorielles discriminantes

Les axes F1 et F2 des AFD réalisées ont un pouvoir discriminant variant de 77.58% à 89.43% pour les quatre jeux de données (Figures 1.15 et 1.16). Une légère démarcation entre les habitats cultivés et non-cultivés est observée pour F2010, C2010 et C2012. Cette tendance n'est en revanche pas observée pour l'échantillonnage F2012.

Plus précisément, pour F2010, on observe une structuration des habitats semi-naturels qui peut être partiellement expliquée par les variables *Tymoviridae*-like, *Nanoviridae*-like, *Partitiviridae*-like et *Geminiviridae*-like (Figure 1.15). D'autre part les milieux cultivés seraient marqués par la présence de *Luteoviridae*-like, ce qui est en accord avec l'analyse SADIE (Figures 1.11 et 1.15). Certains points d'échantillonnages des cultures diversifiées et prairies de F2012 semblent structurés par les *Betaflexiviridae*-like et les *Luteoviridae*-like. Concernant l'analyse de la distribution des variables et des observations pour C2010, on observe que certains points d'échantillonnage en milieux anthropisés sont structurés par trois familles virales : les *Endornaviridae*-like (confirmé par l'analyse SADIE, Figure 1.13), les *Luteoviridae*-like et les *Closteroviridae*-like (Figure 1.16). Enfin, l'échantillonnage C2012 montre que les milieux cultivés sont plutôt structurés par les *Endornaviridae*-like (confirmé par l'analyse SADIE, Figure 1.13) et les *Closteroviridae*-like. On remarque également le regroupement de la majorité des points représentant les habitats non-cultivés qui serait expliqué par la présence de *Caulimoviridae*-like et de *Partitiviridae*–like (Figure 1.16).

Globalement, on remarquera que pour chaque période d'échantillonnage, le milieu non-cultivé n'est pas particulièrement structuré par une seule famille virale mais plutôt par l'addition de plusieurs familles, et *a contrario*, que le milieu cultivé est quant à lui structuré par quelques familles virales uniquement.

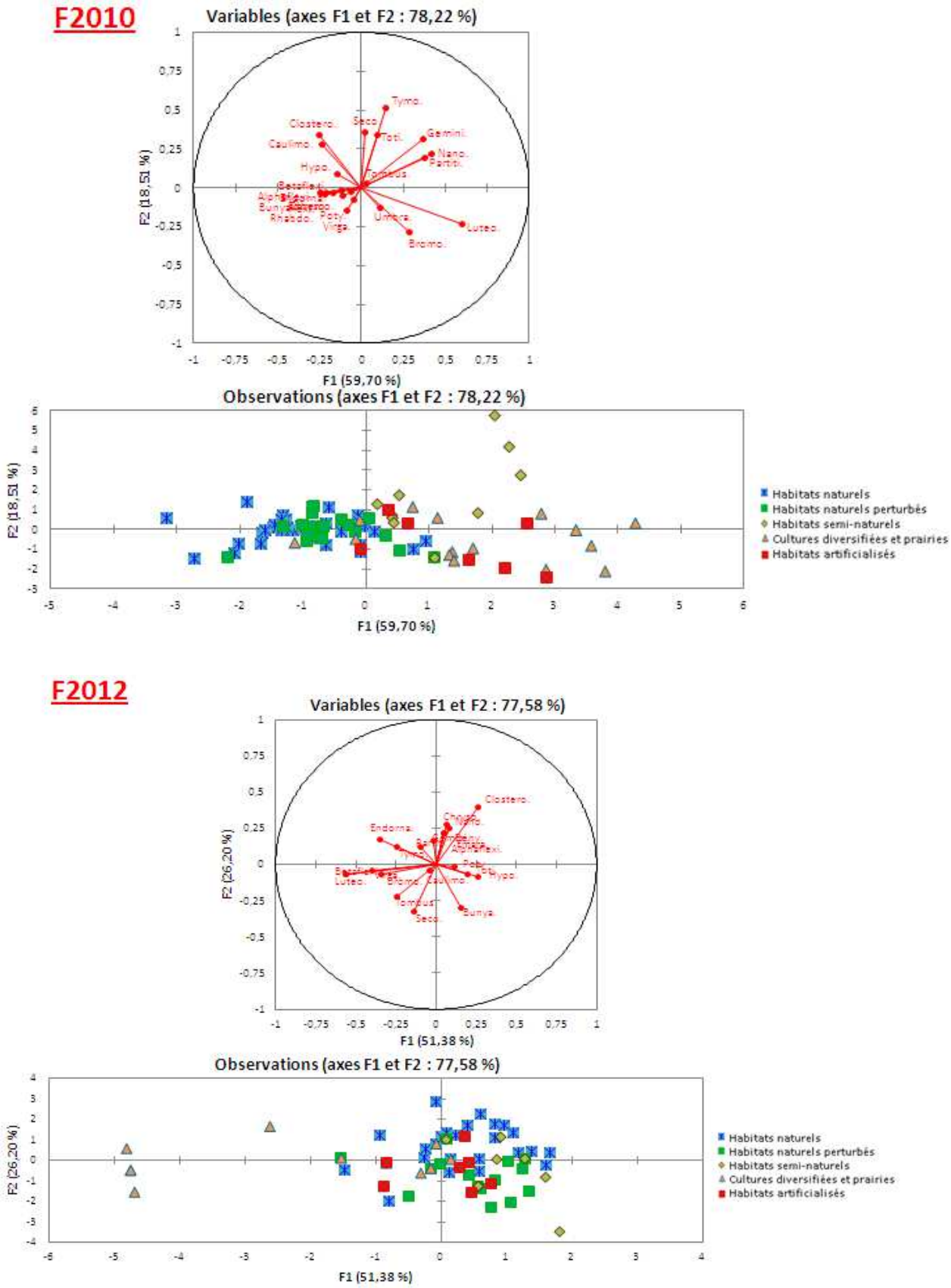


Figure 1.15 : Analyses factorielles discriminantes réalisées sur les échantillonnages de F2010 et F2012. Les points d'échantillonnages ont été utilisés en tant qu'observations, les types d'habitats en tant que variable qualitative et les familles virales en tant que variables quantitatives. Les suffixes « *viridae* » et « virus » ont été supprimés afin d'améliorer la lisibilité.

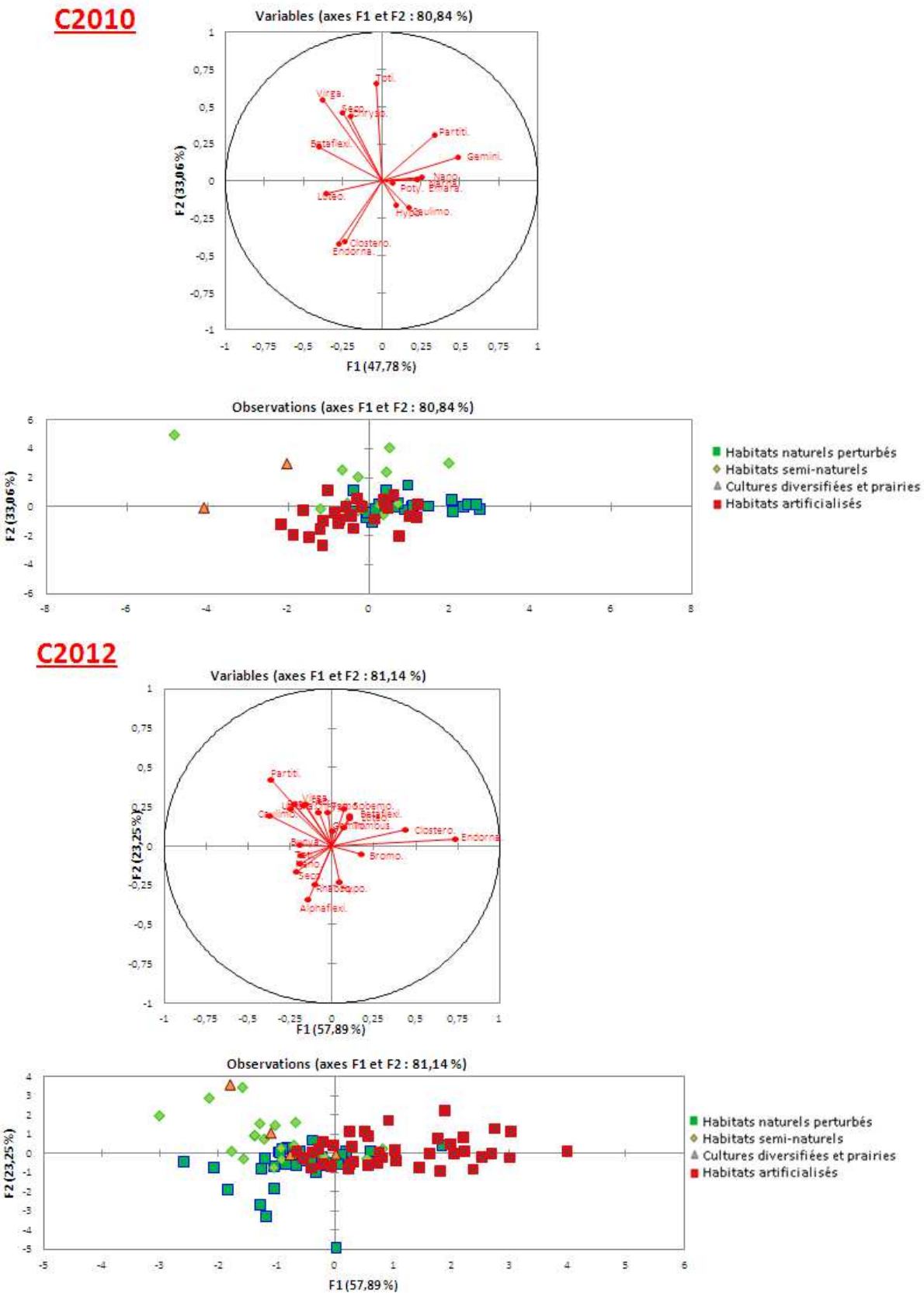


Figure 1. 16 : Analyses factorielles discriminantes réalisées sur les échantillonnages de C2010 et C2012. Les points d'échantillonnages ont été utilisés en tant qu'observations, les types d'habitats en tant que variable qualitative et les familles virales en tant que variables quantitatives. Les suffixes « *viridae* » et « virus » ont été supprimés afin d'améliorer la lisibilité.

3.5.6. Analyse phylogénétique de 9 *Geminiviridae*-like du Fynbos en 2010

Comme reporté dans le Matériel et Méthodes, tous les reads pour lesquels nous avions obtenu un résultat significatif de Blast_N/X (bases de données virales, eValue : 0.001 et identité : 50%) ont été filtrés en fonction des familles virales inféodées aux végétaux mais aussi aux champignons inventoriées par l'ICTV. Nous n'avons pas intégré à cette analyse les mycovirus et d'autres virus proches de geminivirus tels que les gemycircularvirus qui sont encore classés comme « unclassified ssDNA viruses». Cependant, de nouveaux mycovirus environnementaux ou isolés au sein de la phyllosphère ont fait l'objet de plusieurs publications pendant le déroulement de ma thèse (Dayaram *et al.*, 2012; Du *et al.*, 2014; Hafez *et al.*, 2014; Yu *et al.*, 2010). Par ailleurs, des virus proches des geminivirus et des mycovirus dont les origines restent encore inconnues ont été découverts dans des échantillons d'insectes ou encore de matière fécale (Rosario *et al.*, 2012; Sikorski *et al.*, 2013). Nous avons donc décidé de séquencer les génomes totaux de tous les *Geminiviridae*-like détectés au sein de l'échantillon F2010 afin de déterminer leurs positions phylogénétiques et leurs possibles positions taxonomiques.

Nous avons pu obtenir les génomes entiers de 9 *Geminiviridae*-like détectés au sein de l'échantillonnage F2010. Les Blasts effectués à partir des génomes entiers (Tableau 1.7a) ou bien à partir des protéines Rep identifiées à partir de ces génomes (Tableau 1.7b) ont été doublés d'une analyse phylogénétique afin de clarifier la taxonomie de ces 9 virus. Ces analyses ont montré que seul un des neuf virus, A14, semble faire partie de la famille des *Geminiviridae*. Les 8 autres virus semblent se rapprocher des nouveaux groupes viraux évoqués précédemment : les gemycircularvirus et les mycovirus.

Si l'on se focalise sur la phylogénie réalisée à partir de la Rep de ces *Geminiviridae*-like issus de F2010 et d'un groupe de Rep représentatif de la diversité des *Geminiviridae*-like actuellement disponible sur la GenBank, nous pouvons confirmer le fait qu'A14 semble appartenir à cette famille et qu'il est très proche d'un géminivirus qui n'a pas encore de genre attribué : le French bean severe leaf curl virus. La caractérisation du géminivirus isolé de la plante A14 a été réalisée et sera traitée dans la deuxième partie de cette thèse. Les échantillons D94 et K90 semblent quant à eux arborer une Rep proche de celles des mycovirus. Concernant les échantillons E80, B90, E87, A99, B89 et D98, ils forment un groupe monophylétique hautement divergent des gemycircularvirus.

On remarquera également que tous ces génomes viraux, mis à part A14, sont issus du milieu cultivé.

a)		**BlastX Génomes entiers**			
Echantillon	Description	Taxonomie	Query cover	E-value	% Identité
B89	capsid protein [Gemycircularvirus 10]	Gemycirculavirus	35%	6E-77	51%
B90	capsid protein [Sclerotinia sclerotiorum hypovirulence associated DNA virus 1]	Mycovirus	38%	1E-112	60%
D98	capsid protein [Gemycircularvirus 10]	Gemycirculavirus	35%	1E-76	51%
E80	capsid protein [Gemycircularvirus 3]	Gemycirculavirus	36%	4E-97	62%
E87	capsid protein [Sclerotinia sclerotiorum hypovirulence associated DNA virus 1]	Mycovirus	36%	8E-122	65%
K90	capsid protein [Gemycircularvirus 3]	Gemycirculavirus	36%	2E-108	64%
A99	capsid protein [Gemycircularvirus 4]	Gemycirculavirus	37%	5E-90	55%
D94	replication-associated protein [Hypericum japonicum associated circular DNA virus]	Mycovirus	51%	4E-142	65%
A14	replication associated protein [French bean severe leaf curl virus]	*Geminiviridae*	31%	1E-122	68%

b)		**BlastP Rep**			
Echantillon	Description	Taxonomie	Query cover	E-value	% Identité
B89	replication associated protein [Gemycircularvirus 7]	Gemycircularvirus	87%	2E-30	39%
B90	replication associated protein [Gemycircularvirus 7]	Gemycircularvirus	93%	4E-31	41%
D98	replication associated protein [Gemycircularvirus 7]	Gemycircularvirus	87%	6E-30	39%
E80	replication-associated protein [Dragonfly-associated circular virus 1]	Gemycircularvirus	97%	2E-27	39%
E87	replication associated protein [Gemycircularvirus 7]	Gemycircularvirus	79%	5E-31	41%
K90	replication-associated protein [Hypericum japonicum associated circular DNA virus]	Mycovirus	78%	2E-141	100%
A99	replication associated protein [Gemycircularvirus 7]	Gemycircularvirus	87%	3E-30	38%
D94	replication-associated protein [Hypericum japonicum associated circular DNA virus]	Mycovirus	68%	3E-63	72%
A14	replication associated protein [French bean severe leaf curl virus]	*Geminiviridae*	61%	2E-117	79%

Tableau 1.7 : Résultat des Blasts réalisés à partir des génomes entiers (a) et de la Rep (b) des *Geminiviridae*-like détectés grâce à la géo-métagénomique.

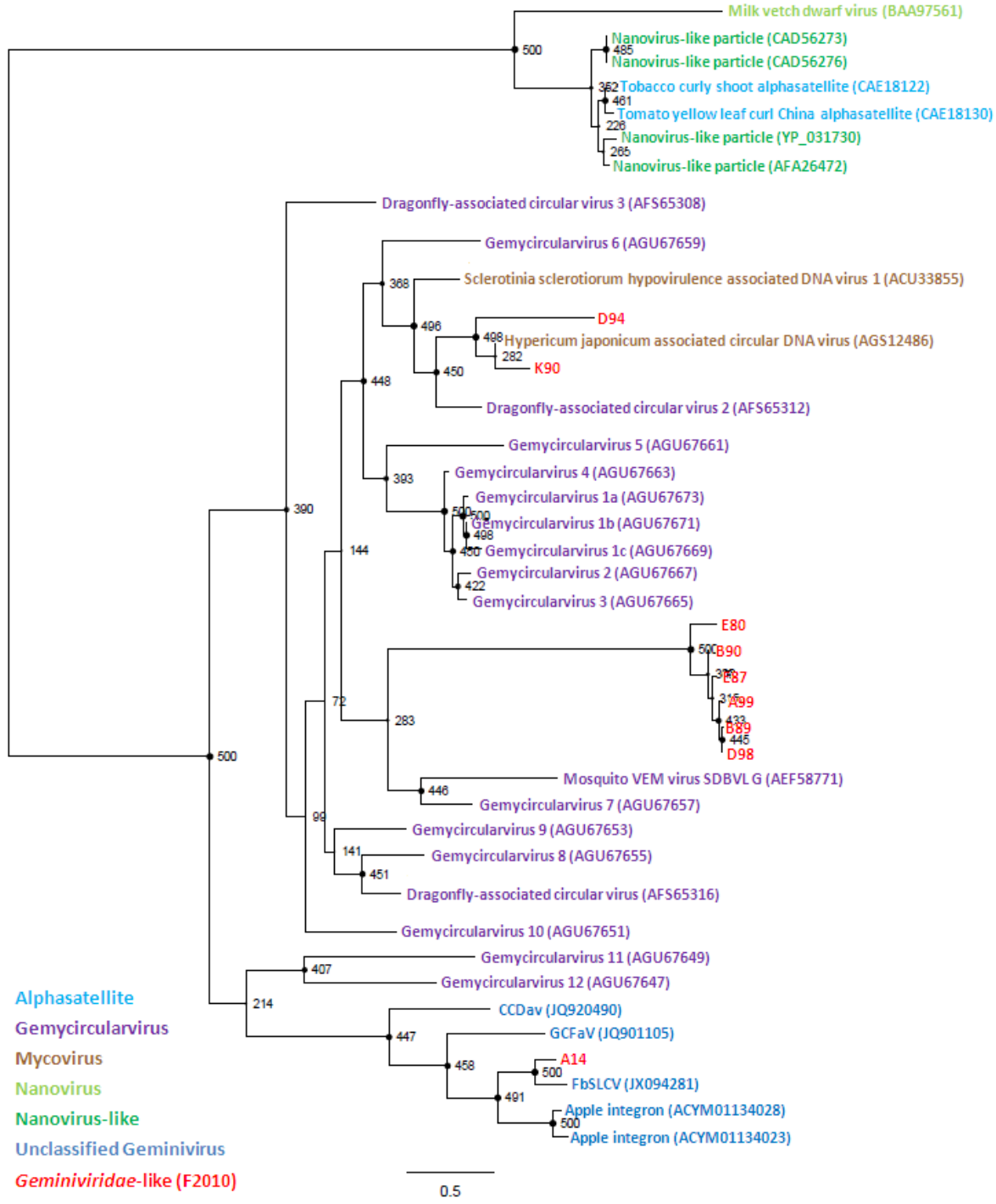


Figure 1.17 : Arbre phylogénétique réalisé selon la méthode de maximum de vraisemblance avec 500 bootstraps selon le modèle évolutif WAG à partir des séquences protéiques de la Rep des 9 *Geminiviridae*-like ainsi que de *Geminiviridae* et autres virus à ssDNA proches. A chaque échantillon est attribuée sa taxonomie par un code couleur. Les bootstraps sont indiqués à la base des nœuds et la taille du nœud est proportionnelle à la valeur de ce bootstrap.

4. Discussion

4.1. Diversité globale des communautés végétales

Les courbes de raréfaction (Figure 1.5) nous ont indiqué que l'effort d'échantillonnage que nous avons fourni au niveau du genre n'était pas assez important de même que celui fourni au niveau de la famille pour les campagnes d'échantillonnage en Camargue. Il est alors important de souligner que cet effort d'échantillonnage a peut-être une influence sur nos résultats, bien que cet effort semble adapté aux échantillonnages au niveau de la famille dans le fynbos. À l'avenir il sera important de réaliser un effort d'échantillonnage plus important, notamment en menant une étude de la diversité végétale au préalable, afin de capturer une représentation réelle de cette diversité dans les zones étudiées.

Nos travaux montrent que les communautés de plantes échantillonnées dans le fynbos sont globalement plus diversifiées que les communautés échantillonnées en Camargue. Ce résultat était attendu, en effet, 8200 espèces végétales ont été recensées dans le fynbos (Myers *et al.*, 2000) contre 1061 en Camargue (http://www.tourduvalat.org/). De plus, le fynbos possède un des taux d'endémisme les plus élevés au monde (68%, http://www.plantzafrica.com/vegetation/fynbos.htm) alors que seulement 7% des plantes de Camargue sont endémiques du bassin méditerranéen Français (http://www.enplr.org/IMG/pdf/Partie_1.1_-_DIAG_SRB-04-04-2008.pdf).

L'indice de Morisita-Horn indique par ailleurs que les communautés de plantes du fynbos sont différentes de celles de Camargue ($0.52<M<0,66$), ce qui était fortement attendu. De façon intéressante, les jeux de plantes échantillonnés en 2010 et 2012 sont très similaires dans chaque région que ce soit en Camargue ou en Afrique du Sud ($0.93<M<0,98$) (Tableau 1.2). Ceci nous indique que les deux écosystèmes sont globalement restés stables sur deux ans et qu'il n'y a pas eu de perturbation majeure évidente qui aurait pu avoir un impact sur les dynamiques virales d'une année sur l'autre.

Nous avons ensuite calculé les indices de diversité de Shannon-Wiener de chaque communauté végétale (au niveau du genre et de la famille) pour chaque habitat (score attribué à l'habitat variant de 0 à 4). On observe globalement une diminution graduelle de la diversité des communautés avec l'augmentation du niveau d'anthropisation du milieu (Figure 1.6). Ceci confirme le fait que l'anthropisation a pour conséquence la diminution de la diversité des communautés végétales. Cependant, comme il est difficile de délimiter clairement les milieux cultivés et non-cultivés, nous nous sommes basés à la fois sur le graphique de la diversité des communautés végétales et sur la définition de chaque type d'habitat pour placer la limite entre milieux cultivés et non-cultivés ; la limite a été placée entre les habitats notés de 0 à 1 pour les milieux non-cultivés et les habitats notés de 2 à 4 pour les milieux cultivés. Comme indiqué ci-dessus, elle, tient

compte de la diversité mais également de l'action de l'Homme sur les communautés botaniques par introduction de plantes cultivées.

4.2. Diversité globale des communautés virales, prévalence de communauté et nombres moyens de « virus » par échantillon de plante virosé

Notre étude montre qu'une très grande majorité des familles de virus de plantes répertoriées par l'ICTV (19/23) ont été identifiées, dont 18/23 en Camargue et 17/23 en Afrique du Sud. Les virus à ARN, majoritairement représentés par les virus à ssRNA+ et dsRNA, sont beaucoup plus fréquents que les virus à ADN sur les deux sites d'étude. L'abondance des virus à ARN dans les écosystèmes étudiés est en cohérence avec leur fréquence relative au sein de la taxonomie des phytovirus (l'ICTV répertorie 19 familles de virus à ARN sur les 23 familles de phytovirus, Annexe 1).

Nos travaux apportent un regard complémentaire aux travaux de Marylin Roossinck (Roossinck, 2012b; Roossinck *et al.*, 2010) dans lesquels 70% des échantillons de plantes sont positifs pour la présence d'au moins un 'virus'. Les pourcentages obtenus dans nos quatre échantillonnages sont tous plus faibles et sont compris entre 26% à 58%. Cependant, il est important de souligner que les écosystèmes étudiés ainsi que la méthode d'extraction des acides nucléiques sont différents entre notre étude et les travaux de Marylin Roossinck (Roossinck, 2012b; Roossinck *et al.*, 2010), ce qui pourrait expliquer cette différence de résultats.

La co-occurrence de virus a été observée dans 8% à 31% de nos échantillons végétaux. Les données sur la propension naturelle des phytovirus à la co-infection sont relativement rares. En effet, la majorité des études concernant les co-infections ont été réalisées sur des interactions hôtes-pathogènes spécifiques (Woolhouse *et al.*, 2002). Toutefois, une étude réalisée *in vitro* à partir de 21 espèces différentes de plantes sauvages et 5 espèces virales à démontré que les virus testés avaient une tendance générale à co-infecter les plantes sauvages étudiées (Malpica *et al.*, 2006). Des co-infections ont ainsi été détectées dans 76% (16/21) des plantes. Ce résultat est difficilement comparable avec celui que nous avons obtenu car la co-occurrence virale tient compte d'une communauté végétale et non d'individus. Il serait donc intéressant de revenir sur chacun de nos échantillons pour lesquels nous avons détecté une co-occurrence virale et d'évaluer la co-infection au sein de chaque individu afin de pouvoir comparer nos résultats à ceux de l'étude citée ci-dessus. Il serait intéressant d'obtenir ce type de résultat car l'étude citée ne prends en compte que 5 espèces virales et 21 espèces de plantes hôtes, elle est probablement loin d'être représentative des agro-écosystèmes naturels tels que ceux sur lesquels nous nous sommes focalisés dans notre étude qui tient compte des diversités de plantes et de virus beaucoup plus importante. Enfin, l'étude citée est forcément biaisée par des conditions expérimentales qui de fait sont incapables de prendre en compte tous les paramètres d'un milieu naturel. Aussi, en l'absence d'études antérieures, notre étude pourrait nous amener à fournir une

première estimation du taux de co-infection naturelle dans deux agro-écosystèmes. L'analyse des indices de Morisita-Horn nous montre que les familles phytovirales ne sont pas structurées géographiquement à l'échelle régionale comme le sont les familles botaniques. Ce résultat semble indiquer que le rang taxonomique de la famille virale n'est pas assez profond dans la taxonomie pour mettre en avant des différences géographiques liées à la spécialisation d'hôte par exemple. Il faudrait donc probablement descendre au niveau du genre, voire de l'espèce pour pouvoir peut être mettre en évidence des phénomènes de différentiation géographique. Nous disposons des données taxonomiques au niveau du genre, mais la synthèse bibliographique et aussi notre propre expérience révèlent que toute tentative d'analyse de diversité à partir de ces données pourrait s'avérer hasardeuse. En effet, nous pouvons citer l'exemple d'un géminivirus découvert sur luzerne pour l'échantillonnage C2010. Les reads concernant ce géminivirus ont été classés dans 3 OTUs différentes au niveau du genre (Begomovirus, Mastrevirus et Geminivirus non-classifié), ce qui nous aurait amené à considérer la présence de 3 genres différents de virus dans la même plante. Or à l'aide d'une analyse plus poussée de séquences génomiques complètes générées à partir des plantes correspondantes (cela sera traité dans le chapitre 2 de cette thèse), ces 3 OTUs appartiennent en réalité à une seule et même espèce virale. Par conséquent, si les calculs de diversité virale avaient été réalisés au niveau du genre, les inventaires viraux auraient été surestimés en raison de très nombreux faux-positifs. En reprenant la formule célèbre de Mr. Bass Becking : « Everything is everywhere, but, the environment selects » (Baas Becking, 1934), on pourrait conclure que la majorité des familles virales sont « partout » mais que par sélection (mais aussi par dérive) les espèces se différencient et ne sont probablement pas réparties partout. Ainsi, de la même manière que les communautés de plantes (à l'échelle de la famille botanique), les espèces virales présentent vraisemblablement des signaux d'adaptation et *de facto* de différentiation géographique. Malheureusement, la technique de pyroséquençage ne permet pas de descendre l'analyse des communautés virales à l'échelle de l'espèce par manque de couverture et de profondeur des reads. Les progrès technologiques tel que ceux qui sont mis en œuvre pour la technique Illumina par exemple, permettront certainement l'assemblage de génomes viraux entiers afin de réaliser ce type d'études à l'échelle spécifique.

4.3. Prévalences virales en relation avec l'agriculture et le statut des plantes

Nos travaux montrent que pour trois des quatre périodes d'échantillonnage (F2010, F2012 et C2012), les prévalences virales de communauté sont significativement plus fortes dans le milieu cultivé que dans le milieu non-cultivé (Figure 1.8 et Tableau 1.5). Pour F2010, l'association type la plus fréquente dans le groupe des plantes virosées est la suivante : « plante cultivée exotique et annuelle ». Pour F2012, on retrouve le même type d'association alors que pour C2012 c'est uniquement « plante cultivée ». Nos travaux semblent donc suggérer que le milieu cultivé est soumis à des épisodes

d'épidémies virales traduits par une augmentation transitoire de la prévalence virale. En revanche, nos travaux montrent une certaine stabilité de la prévalence virale en milieu non-cultivé. De façon intéressante, pour C2010, seule collecte pour laquelle une tendance épidémique n'a pas été répertoriée, la prévalence des plantes adventices virosées du milieu cultivé était significativement plus importante que celle des plantes cultivées. Ce résultat suggère que les adventices présentent probablement des types d'interaction avec les virus différents de ceux observés chez leurs proches voisins cultivés.

Nos résultats sont cohérents avec le fait que le milieu cultivé est généralement considéré comme plus sensible aux agents pathogènes (Edwards, 1996; Elena *et al.*, 2014), et que le phénomène de course aux armements, exacerbé par les interventions humaines, y induirait une forte augmentation de la virulence virale (Bergelson *et al.*, 2001; Brown and Tellier, 2011). Nos résultats sont par ailleurs confortés par une étude réalisée sur le piment sauvage au Mexique qui a démontré que la prévalence virale augmentait avec le niveau d'anthropisation (Pagan *et al.*, 2012; Rodelo-Urrego *et al.*, 2013). De plus, on peut ici invoquer une des hypothèses émises quant à la relation entre la diversité végétale et le risque d'épidémie qui est celle de l'effet de dilution : plus la diversité végétale est forte, plus le risque d'épidémie est faible, et vice versa (Keesing *et al.*, 2010; Keesing *et al.*, 2006). Par ailleurs, les plantes du milieu sauvage auraient co-évolué avec leurs pathogènes associés (Lovisolo *et al.*, 2003), ce qui aurait amené à un équilibre et une stabilité (« guerre des tranchées »,), ce qui pourrait expliquer le fait que les épidémies y soient plus rares.

Nos travaux indiquent par ailleurs que les prévalences virales dans les échantillons de plantes exotiques d'Afrique du Sud sont significativement plus élevées que celles des échantillons de plantes indigènes quelle que soit l'année de prélèvement. On observe ainsi que les échantillons de plantes exotiques manifestent des prévalences virales de 9% (F2012) à 24% (F2010) plus importantes que celles des échantillons de plantes indigènes. Ce résultat va dans le sens de l'hypothèse du « pathogen spill-over » qui théorise que les plantes exotiques pourraient être un lieu d'augmentation des concentrations virales menaçant ainsi les plantes indigènes (Power and Mitchell, 2004). Cette hypothèse a été testée et validée sur le pathosystème *Avena fatua* / B/CYDVen Californie. Il a été montré dans ces travaux qu'*Avena fatua*, qui est une graminée exotique tolérante au B/CYDV, est responsable de l'augmentation de la prévalence de ces virus dans les communautés de graminées natives. Le fait que les communautés natives aient été plus impactées par ces virus qu'*A. fatua* a permis par ailleurs à cette espèce introduite d'envahir la communauté (Malmstrom *et al.*, 2005a; Malmstrom *et al.*, 2005b; Power and Mitchell, 2004). Ce type d'expérience a permis de démontrer le potentiel d'*A. fatua* en tant que réservoir viral participant au pathogen spillover (Power and Mitchell, 2004). Même si notre étude de géo-métagénomique ne permet pas de tester directement ce genre d'hypothèse, elle permet tout au moins de la renforcer en montrant que les plantes exotiques peuvent potentiellement jouer le rôle de réservoirs viraux avec un risque pour les plantes indigènes. A partir de notre jeu de données, il

serait possible de cibler des plantes d'intérêt et de tester à l'échelle de l'agro-écosystème la théorie du « pathogen spillover » sur la base des travaux d'Alison G. Power et Charles E. Mitchell.

4.4. Répartition spatio-temporelle des phytovirus et analyses factorielles discriminantes

Les analyses combinées de la répartition spatiotemporelle globale des familles virales et des AFD ont révélé une tendance générale à l'absence de patron de répartition des familles virales à l'échelle des écosystèmes français et sud-africains. Ce constat renforce l'idée développée précédemment que les familles virales sont généralement réparties de manière ubiquitaire.

Toutefois il s'avère que certaines familles présentent des patrons de distribution spatiale particuliers. Les résultats des deux types d'analyses ont notamment montré que les *Endornaviridae*-like de Camargue étaient inféodés au milieu cultivé que ce soit en 2010 ou en 2012. Les plants « infectés » sont majoritairement du riz. Or, comme le riz est connu pour transmettre les *Endornaviridae* de façon verticale, on peut comprendre que les patrons de leur distribution ne soient pas conservés d'une année sur l'autre. En effet, le riz étant une plante annuelle, la distribution spatiale serait chaque année le reflet du semis aléatoire de graines de riz porteuses de virus.

Les analyses spatio-temporelles semblent également indiquer que les *Luteoviridae*-like sont des virus que l'on retrouve avec des prévalences plus fortes dans le milieu cultivé. Ce résultat n'est pas étonnant dans le sens où la famille *Luteoviridae* contient diverses espèces connues pour leurs effets dévastateurs sur les cultures(Astier, 2007; Hull, 2002). De manière assez étonnante, les *Closteroviridae*-like (qui représentent une forte proportion des virus dans chacune de nos campagnes d'échantillonnage) sont plutôt inféodés au milieu cultivé en Camargue et au milieu non-cultivé dans le fynbos. Ce résultat est cependant cohérent avec le fait que ces virus sont connus pour avoir une large gamme d'hôte (http://pvo.bio-mirror.cn/) - ils peuvent infecter des monocotylédones mais aussi des eudicotylédones - et être transmis par insectes vecteurs.

Les AFD que nous avons réalisées montrent que pour chaque période d'échantillonnage, le milieu non-cultivé est structuré par un large ensemble de variables mineures (familles virales) alors que le milieu cultivé est quant à lui structuré par quelques familles virales. Ce résultat semble à nouveau montrer que le milieu non-cultivé héberge une diversité virale plus importante que le milieu cultivé et que les épidémies en milieu sauvage sont plus rares comme cela a déjà été suggéré dans la revue de Ian Cooper et Roger A.C. Jones (Cooper and Jones, 2006). A contrario, ce résultat semble traduire le fait que le milieu cultivé favorise l'apparition d'épidémies et par conséquence l'augmentation de la prévalence de quelques virus (*Endornaviridae*-like, *Luteoviridae*-like).

Un patron que nous avons jugé intéressant d'aborder est celui de la distribution spatio-temporelle des *Partitiviridae*-like. En effet, il s'avère que quelque soit la période et le lieu d'échantillonnage, ces virus ont une distribution conservée qui s'applique généralement aux plantes non-cultivées (sauf pour F2010). Cependant, des variations de prévalence et de distribution de cette famille virale ont été observées dans nos travaux. Ces variations peuvent être dues à des erreurs techniques (les plantes échantillonnées en 2012 ne seraient pas issues de la descendance de celles échantillonnées en 2010, problèmes de conservation des échantillons, etc.) ou à des phénomènes biologiques tels que l'assainissement des plants malades dans le fynbos entre 2010 et 2012 ou à l'inverse d'une propagation de la descendance infectée en Camargue entre 2010 et 2012. On soulignera par ailleurs que Marylin Roossinck, lors de son étude d'écogénomique, a démontré que les *Partitiviridae* étaient présents majoritairement dans les plantes échantillonnées alors que ce n'est clairement pas le cas pour nos échantillons (Figure 1.7). En Afrique du Sud, le fait que ces virus soient présents de façon conservée uniquement au niveau de plantes sauvages situées dans des zones 0 ou 1 mais ne le sont pas dans des zones cultivées renvoie à deux idées développées par M . Roossinck : (i) que les *Partitiviridae* ont un mode de vie persistant et que leur transmission se fait via la graine (Roossinck, 2010; Roossinck, 2012a) et (ii) que certains virus persistants pourraient être associés à des interactions de type mutualiste avec leur hôte (Roossinck, 2005; Wren *et al.,* 2006). La présence ou surtout l'absence de cette famille de virus pourrait ainsi être une sorte de marqueur de la « naturalité » de l'écosystème et par la même du niveau de perturbation humaine subit par le milieu.

4.5. Apport de l'obtention de génomes entiers

Nos travaux menés sur des génomes entiers de virus, classés comme appartenant au groupe *Geminiviridae*-like nous a tout d'abord permis de valider l'approche développée dans notre étude de métagénomique, à savoir de choisir un read comme seuil de détection d'un virus. Nous avons ainsi montré que la présence d'un seul read pouvait amener à la détection de la présence d'un virus (Tableau 1.8). Par exemple, le read unique issu d'une euphorbe sauvage (Echantillon A14, *Euphorbia caput-medusae*) et attribué à la famille *Geminiviridae* était la trace effective de la présence d'un nouveau géminivirus (voir chapitre II). Les 8 autres génomes du groupe *Geminiviridae*-like étaient quant à eux associés à un nombre faible et variable de reads allant de 1 à 35 (Figure 1.8). Par ailleurs, l'obtention des génomes complets nous a montré que la grande majorité de ces virus n'appartiennent pas à la famille *Geminiviridae* mais probablement à deux nouveaux genres viraux encore non-répertoriés par l'ICTV : les mycovirus et les gemycircularvirus. Les hôtes de ces virus à ssDNA ne sont pas encore clairement identifiés mais il semble qu'une grande majorité d'entres eux pourraient être inféodés aux champignons (Dayaram *et al.*, 2012; Hafez *et al.*, 2014; Liu *et al.*, 2012a; Rosario *et al.*, 2012; Sikorski *et al.*, 2013; Yu *et al.*, 2010). La fréquence avec laquelle ces virus ont été trouvés parmi les virus à ADNs simple brin pourrait suggérer qu'ils « passent » dans les vaisseaux végétaux, une hypothèse séduisante qui mérité d'être testé.

Echantillon	Nombre de reads 454
B89	1
B90	15
D98	4
E80	35
E87	1
K90	1
A99	2
D94	16
A14	1

Tableau 1.8 : Nombre de reads correspondant à des *Geminiviridae*-like en fonction des échantillons dans lesquels ils ont été détectés.

Nos travaux de séquençage de génomes complets à partir de leurs traces métagénomiques représentent un cas d'étude. Ils soulignent l'intérêt de l'obtention de génomes entiers pour mieux caractériser les paramètres écologiques influençant les dynamiques virales au sein des écosystèmes. Une vision plus précise de la diversité au niveau du genre et de l'espèce devrait permettre d'affiner les patrons de répartition spatiale et de discriminer finement la diversité phytovirale à l'échelle de l'agro-écosytème. Les nouvelles techniques de séquençage comme Illumina MySeq (2 fois 300 pb ou plus) ou Pacific Bioscience pourront peut être dans un proche avenir permettre d'atteindre cet objectif.

Beaucoup d'autres nouveau virus, révélés pour la première fois dans notre étude, mériteraient d'être caractérisés par un séquençage complet de leur génome. Cependant, par manque de moyens et de temps, ce travail plus exhaustif n'a pas pu être réalisé au cours de la thèse car il concerne potentiellement 411 nouvelles espèces virales et 189 nouveaux variants/isolats viraux (Annexes 5 et 6). Une des perspectives majeures d'exploitation de cette mine restera donc à déterminer l'identité taxonomique des nouveaux virus révélés ici pour étendre la connaissance de la sphère phytovirale mondiale.

5. Conclusion générale

Les travaux présentés dans cette première partie de thèse avaient pour but de répondre à trois questions. Tout d'abord nous voulions tester si le milieu sauvage représente un réservoir de biodiversité virale. Les résultats attendus pour confirmer cette hypothèse étaient de trouver une diversité plus forte dans le milieu non-cultivé. Cependant, notre étude a révélé que l'approche employée, la géo-métagénomique basée sur du pyroséquençage 454, ne permet de calculer que des indices de diversité au niveau des familles virales. A ce niveau taxonomique, les diversités en milieu cultivé et non-cultivé ne sont pas significativement différentes quels que soient les lieux ou les années d'échantillonnage. Les résultats seront probablement différents si la profondeur et la couverture des génomes viraux pouvaient être augmentées ce qui sera

probablement réalisable avec les nouvelles générations de séquençage à haut-débit (Pacific Bioscience ou Illumina).

Nous nous étions également attachés à évaluer l'influence de la diversité végétale sur les prévalences virales de communauté. Notre étude de géo-métagénomique a montré que le milieu cultivé est significativement associé à des prévalences virales de communauté plus élevées que le milieu non-cultivé dans la majorité des cas (3/4 campagnes d'échantillonnage). Ce résultat est hautement important car il apporte un support expérimental à une hypothèse classique abondamment invoquée mais rarement testée à une grande échelle *in situ*. Cet état de fait peut être expliqué, côté milieu naturel par un effet de dilution, et côté agricole par la fragilité des cultures liée à une perte de diversité botanique, à une fragilisation des cultivars à haut rendement, à un manque de co-évolution sur le long terme avec les virus.

Par ailleurs, nous avons pu identifier certains paramètres écologiques impliqués dans les variations de prévalences à l'échelle de l'agro-écosystème comme le statut de la plante (exotique vs. indigène ou encore cultivée vs. non cultivée). Nous confirmons dans nos travaux à très grande échelle que les échantillons plantes exotiques (souvent envahissantes) ont des prévalences virales plus importantes que les échantillons plantes indigènes. Ce résultat est très important car il rend crédible l'hypothèse du « Pathogen spillover », qui stipule qu'une épidémie observée sur une population-hôte peut être provoquée non pas par la transmission du pathogène à l'intérieur de la population-hôte mais par la transmission de l'agent pathogène à partir d'une population réservoir, à l'échelle de l'agro-ecosystème. Il serait intéressant de tester dans le fynbos cette hypothèse à partir de plusieurs plantes envahissantes (*Briza maxima*, *Avena fatua*, *Bromus diandrus* etc.).

Enfin, nous avons pu montrer que certaines familles virales (mais pas la majorité) présentent des patrons en terme de distribution spatio-temporelle qui peuvent notamment être expliqués par leur mode de transmission, l'intervention humaine ou encore leur gamme d'hôte. Même si la fréquence de nombreux virus était manifestement trop faible pour en tirer des informations spatiales et temporelles, notre étude marque un point de départ dans ce domaine qui reste encore à explorer pour la majorité des virus.

Sur la base de ces premiers résultats, la géo-métagénomique peut être considéré comme un outil incontournable pour (i) étudier les prévalences virales de communauté dans les écosystèmes et les facteurs qui interviennent dans la structuration de ces prévalences et (ii) inventorier et découvrir de nouveaux virus présents dans des environnements encore largement inexplorés.

Chapitre II : Découverte et caractérisation d'un nouveau genre appartenant à la famille *Geminiviridae.*

1. Contexte et objectifs

Grâce aux études de géo-métagénomique menées dans le fynbos et en Camargue, nous avons pu détecter des séquences correspondant à 3 nouveaux virus relativement proches en terme de séquences les uns des autres au sein de la famille *Geminiviridae*. La géo-métagénomique nous a permis de retracer l'identité des hôtes sur lesquels ces geminivirus ont été détectés. Le premier a été détecté en 2010 sur une euphorbe sauvage du fynbos, *Euphorbia caput-medusae* et les deux autres ont été découverts sur des plants de luzerne cultivée, *Medicago sativa*, en Camargue en 2010 et 2012. La particularité de ces 3 geminivirus est que les reads et contigs obtenus se sont révélés être phylogénétiquement très divergents des sept genres établis par l'ICTV au sein de la famille *Geminiviridae* : Begomovirus, Mastrevirus, Curtovirus, Topocuvirus, Becurtovirus, Turncurtovirus et Eragrovirus (Varsani *et al.*, 2014b). De plus, lors de la dernière année de ma thèse, une étude de métagénomique (collaboration BGPI – Université d'Helsinki) visant à décrire la diversité phytovirale du plantain (*Plantago lanceolata*) sur l'archipel d'Åland en Finlande a permis de détecter un autre geminivirus proche de ceux que nous avons découverts en France et en Afrique du Sud.

Le génome des geminivirus est composé de molécules d'ADN circulaires simple brin. Il est composé d'une seule molécule d'ADN pour les geminivirus monopartite ou de deux molécules pour les geminivirus bipartite. Ces molécules d'ADN circulaires comptent moins de 3640nt (Loconsole *et al.*, 2012) et sont encapsidées dans des particules isocaédriques jumelées (Jeske, 2009). Certains membres du genre Begomovirus peuvent être accompagnés de molécules d'ADN satellites dont ils sont plus ou moins dépendants pour la réalisation de leur cycle biologique (Jeske, 2009). Outre les virus qui ont été officiellement classé dans les 7 genres de la famille *Geminiviridae*, les bases de données internationales contiennent des séquences relativement divergentes de geminivirus qui n'ont pu être assignés à aucun des genres existants. La définition du genre chez les geminivirus a été longtemps basée sur leur gamme d'hôte, leur vecteur, l'organisation de leur génome et leur taux de similarité de séquences (Jeske, 2009). Ainsi l'un des critères pour la définition des begomovirus a été leur transmission par aleurode, pour les curtovirus, les mastrevirus, les turncurtovirus et les becurtovirus, leur transmission par cicadelles et pour les topocuvirus, leur transmission par membracides (Varsani *et al.*, 2014b). Avec le déferlement de nouvelles séquences dont plusieurs étaient inclassables dans les genres existants, un vif débat a été engagé au sein du groupe *Geminiviridae* de l'ICTV. pour examiner la pertinence de définir une espèce et un genre sur la seule base de l'identité de séquences (Varsani *et al.*, 2014a). En effet, le genre Eragrovirus a été établi sans avoir connaissance de son vecteur (Varsani *et al.*, 2014b). En ce qui concerne leur organisation génomique, deux gènes sont conservés sur l'ensemble des virus de la famille *Geminiviridae*: le gène *rep* codant pour la protéine associée à la réplication (Rep) et le gène *cp* codant pour la protéine de capside (CP). Une autre constante des géminivirus est leur grande région intergénique (large intergenic région, LIR) qui entoure l'origine de réplication du virion et qui contient les séquences

promotrice pour les gènes du brin viral et ceux du brin complémentaire (Varsani *et al.*, 2014b). Outre ces points communs, chaque genre présente des caractéristiques génomiques spécifiques qui le définissent. A titre d'exemple, les mastrevirus possèdent une Rep épissée et une région intergénique courte (short intergenic région, SIR) en plus de la LIR (Jeske, 2009). Les geminivirus ont une forte propension à la recombinaison ce qui a généré de nouvelles espèces mais également de nouveaux genres (Briddon *et al.*, 1996; Padidam *et al.*, 1999; Stanley *et al.,* 1986; Varsani *et al.*, 2009). En outre, les géminivirus présentent des taux de substitutions remarquablement élevés pour des virus qui dépendent des polymérases de l'hôte pour leur réplication ; leur taux de substitution est, comparable à celui des virus à ARN (Lefeuvre *et al.*, 2007b).

Un grand nombre d'espèces virales appartenant aux *Geminiviridae* a été associé ces dernières décennies à des épidémies affectant des plantes d'intérêt économique telles que la tomate, le coton, le manioc, ou encore le maïs (Moffat, 1999; Rey *et al.*, 2012; Rybicki and Pietersen, 1999). Il a été suggéré que certains de ces géminivirus tels que la souche A du *Maize streak virus* (MSV-A) auraient émergés à partir de plantes sauvages (Monjane *et al.*, 2011). Même si la diversité des geminivirus inféodés aux plantes sauvages reste globalement méconnue, quelques travaux récents se sont traduits par quelques résultats : (i) la détection d'une nouvelle espèce de géminivirus sur la plante sauvage *Eragrostis curvala,* ce qui a permis de décrire un nouveau genre - Eragrovirus (Varsani *et al.*, 2009) et (ii) la découverte de nouvelles espèces capables d'infecter des plantes sauvages (Garcia-Andres *et al.*, 2006; OOi *et al.*, 1997; Varsani *et al.*, 2009)

Dans cette deuxième partie de thèse, le but sera donc de caractériser les génomes complets des trois nouveaux geminivirus potentiels détectés par géo-métagénomique en utilisant des techniques classiques de biologie moléculaire (RCA, PCR etc.) et de séquençage (Sanger). Nous tenterons alors d'élucider le positionnement phylogénétique de ces nouveaux isolats au sein des *Geminiviridae* et de rechercher des traces de recombinaison. Nous chercherons par ailleurs à estimer la prévalence de ces nouveaux géminivirus à l'intérieur et à l'extérieur des zones d'échantillonnage d'origine. Enfin, nous mettrons en place des travaux visant à connaître leurs gammes d'hôte et leurs voies de dissémination (vection). L'objectif final est de gagner des informations sur l'histoire évolutive des *Geminiviridae* et de statuer sur l'émergence potentielle d'un nouveau genre au sein de cette famille de virus. Ce deuxième chapitre de la thèse sera divisée en trois parties: (i) nous ferons une caractérisation moléculaire et biologique du nouveau genre auquel appartiennent ces virus, (ii) nous étudierons la diversité et la prévalence des espèces découvertes sur euphorbe et sur luzerne, et (iii) nous essaierons de caractériser le mode de vection de ce nouveau genre de *Geminiviridae*.

2. Les capulavirus : description d'un nouveau genre de geminivirus divergent et implications taxonomiques

Cette partie a pour but de décrire le genre Capulavirus découvert au travers des études de métagénomique, en caractérisant les différents génomes complets obtenus grâce à leur organisation génomique, la détection d'évènement de recombinaison, et leur position phylogénétique. Pour aller plus loin nous allons à partir de nos résultats inférer des hypothèses sur l'histoire évolutive des *Geminiviridae.* Enfin nous essaierons de comprendre comment un genre viral appartenant à une famille qui est minutieusement scrutée depuis plus de 20 ans par une communauté de virologues extrêmement active dans beaucoup de pays de monde ait pu passer inaperçu jusqu'à récemment. Il est en effet surprenant que plusieurs membres de ce nouveau genre aient été découverts sur une période relativement courte (3 ans), en différents endroits du monde, à la fois sur plantes sauvages et cultivées.

2.1. Identification et caractérisation d'*Euphorbia caput-medusae latent virus*, implication taxonomique et reconsidération de l'histoire évolutive des *Geminiviridae*

L'étude concernant cette partie de la thèse est présentée dans l'article ci-après publié dans Virus Research en juillet 2013.

Virus Research 177 (2013) 35–45

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Identification and characterisation of a highly divergent geminivirus: Evolutionary and taxonomic implications

Pauline Bernardo[a,b], Michael Golden[c], Mohammad Akram[d], Naimuddin[d], Nagaswamy Nadarajan[e], Emmanuel Fernandez[a], Martine Granier[a], Anthony G. Rebelo[f], Michel Peterschmitt[a], Darren P. Martin[c], Philippe Roumagnac[a,*]

[a] CIRAD/UMR BGPI, TA A54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France
[b] INRA/UMR, BGPI, TA A54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France
[c] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, UCT Faculty of Health Sciences, Observatory 7925, South Africa
[d] Division of Crop Protection, Indian Institute of Pulses Research, Kalyanpur, Kanpur 208024, India
[e] Indian Institute of Pulses Research, Kalyanpur, Kanpur 208024, India
[f] South African National Biodiversity Institute, Kirstenbosch, Private Bag X7, Claremont, 7735 Cape Town, South Africa

ARTICLE INFO

Article history:
Received 8 April 2013
Received in revised form 8 July 2013
Accepted 9 July 2013
Available online 22 July 2013

Keywords:
DNA viruses
Geminivirus
Plant viruses
Phylogenetic analysis
Evolution
Taxonomy

ABSTRACT

During a large scale "*non a priori*" survey in 2010 of South African plant-infecting single stranded DNA viruses, a highly divergent geminivirus genome was isolated from a wild spurge, *Euphorbia caput-medusae*. In addition to being infectious in *E. caput-medusae*, the cloned viral genome was also infectious in tomato and *Nicotiana benthamiana*. The virus, named Euphorbia caput-medusae latent virus (EcmLV) due to the absence of infection symptoms displayed by its natural host, caused severe symptoms in both tomato and *N. benthamiana*. The genome organisation of EcmLV is unique amongst geminiviruses and it likely expresses at least two proteins without any detectable homologues within public sequence databases. Although clearly a geminivirus, EcmLV is so divergent that we propose its placement within a new genus that we have tentatively named Capulavirus. Using a set of highly divergent geminiviruses genomes, it is apparent that recombination has likely been a primary process in the genus-level diversification of geminiviruses. It is also demonstrated how this insight, taken together with phylogenetic analyses of predicted coat protein and replication associated protein (Rep) amino acid sequences indicate that the most recent common ancestor of the geminiviruses was likely a dicot-infecting virus that, like modern day mastreviruses and becurtoviruses, expressed its Rep from a spliced complementary strand transcript.

1. Introduction

Among plants viruses, those of the family *Geminiviridae* are responsible for a disproportionately large number of recently emergent crop diseases worldwide. They have dramatically impacted agricultural yields over the past 50 years (Moffat, 1999), and are a major threat to the food security of developing countries in the tropical and sub-tropical regions of the world (Rey et al., 2012; Rybicki and Pietersen, 1999). Most at risk are countries in sub-Saharan Africa where reports near the beginning of the 1900s of diseases in exotic introduced cultivated staple food species such as cassava and maize provided the first clear descriptions of geminivirus infections (Fuller, 1901; Warburg, 1894). Caused by at least seven different African geminivirus species, cassava mosaic disease (CMD) is today recognised as the most important biotic constraint of cassava production throughout this region (Legg and Fauquet, 2004; Patil and Fauquet, 2009). For instance, a recent CMD epidemic affected at least nine countries in East and Central Africa (spanning an area of 2.6 million square kilometres) inflicting annual economic losses of US$1.9–2.7 billion (Patil and Fauquet, 2009). Similarly, throughout sub-Saharan Africa the geminivirus species that causes maize streak disease (MSD) inflicts annual losses averaging approximately US$120–480 million (Martin and Shepherd, 2009). In addition, a range of other African geminivirus species have been described in the past three decades that, while obviously causing serious yield reductions in tomatoes, beans, and sweet-potatoes, have a currently unquantified impact on African agriculture (Rey et al., 2012).

* Corresponding author. Tel.: +33 (0)499 62 48 58.
E-mail address: philippe.roumagnac@cirad.fr (P. Roumagnac).

0168-1702 © 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-SA license.
http://dx.doi.org/10.1016/j.virusres.2013.07.006

All characterised geminivirus genomes are composed of one or two circular single stranded DNA components each of which contains fewer than 3640 nucleotides (Loconsole et al., 2012). Each genomic component is encapsidated in a geminate (twinned) incomplete icosahedral particle. To date, seven genera have been approved within the *Geminiviridae* family: Begomovirus, Curtovirus, Topocuvirus, Mastrevirus, Becurtovirus, Turncurtovirus, and Eragrovirus (http://talk.ictvonline.org/files/ictv_official_taxonomy_updates_since_the_8th_report/m/plant-official/4454.aspx). The criteria for demarcating genera in the family *Geminiviridae* are genome organisation, insect vector, host range and sequence relatedness (Fauquet and Stanley, 2003). It is noteworthy, however, that not all of these criteria are necessary for the approval of new genera as, for example, the genus Topocuvirus has been distinguished based only on sequence relatedeness and vector species while the genus Mastrevirus includes viruses that have multiple different vector species and infect either monocotyledonous or dicotyledonous hosts. Mastrevirus genomes also have the fewest genes and these viruses express only four different proteins with two of these, the replication associated protein (Rep) and RepA, sharing identical N-termini but distinct C-termini (they are expressed from alternatively spliced versions of the same transcript). By contrast the other six genera have between 5 and 8 genes, with only the coat protein (*cp*) and *rep* genes being detectably homologous across all of the genera.

The high diversity of geminivirus genome sequences is likely facilitated by these viruses having much higher mutation and recombination rates than those seen in many other DNA viruses. Despite geminiviruses utilising host DNA polymerases during their replication, their mutation rates are as high as many RNA viruses that replicate using error prone RNA dependent RNA polymerases (Duffy and Holmes, 2008; Ge et al., 2007; Isnard et al., 1998). Whereas it is most likely that the high recombination rates of geminiviruses occur as a consequence of their replication involving a mixture of rolling circle and recombination dependent mechanisms (Jeske, 2009), the generally broad host ranges of these viruses together with the frequent occurrence in nature of mixed infections (Martin et al., 2011) has resulted in both frequent instances of inter-species recombination (Padidam et al., 1999), and occasional instances of inter-genus recombination (Briddon et al., 1996; Stanley et al., 1986). While recombination events have sometimes yielded new geminivirus species, it is also plausible that past inter-genus recombination events may have yielded new geminivirus genera (Briddon et al., 1996; Stanley et al., 1986).

The development and application of rolling circle amplification (RCA) based approaches to discover novel circular ssDNA viruses from a variety of environmental sources (Delwart, 2012; Ng et al., 2011b) has tremendously accelerated the rate at which such viruses have been discovered (Rosario et al., 2012) and, when applied to the study of plant samples, has revealed that geminivirus diversity likely far exceeds that which is currently known (Haible et al., 2006; Ng et al., 2011a; Schubert et al., 2007). Besides the characterisation of divergent curtoviruses, mastreviruses and begomoviruses, various geminivirus species have been discovered that are so divergent that they cannot be convincingly assigned to any of the four established geminivirus genera (Briddon et al., 2010; Loconsole et al., 2012; Varsani et al., 2009; Yazdi et al., 2008). Both the creation of new genera such as Becurtovirus, Eragrovirus and Turncurtovirus to accommodate some of these divergent species and perpetually growing numbers of new species within the existing "older" genera underline the steadily increasing complexity of geminivirus taxonomy and the need to recurrently re-evaluate the objectivity and meaningfulness of genus and species demarcation criteria that are applied to the members of this family (Muhire et al., 2013).

Amongst the most divergent of these newly discovered geminiviruses is Eragrostis curvula streak virus (ECSV) isolated from an uncultivated African grass species, *Eragrostis curvula* (Varsani et al., 2009). Given both that there exists a tremendous bias favouring the discovery of novel viruses in cultivated species, and that many novel geminivirus species have in the past been discovered in uncultivated hosts (Briddon et al., 2010; Tan et al., 1995; Varsani et al., 2009), it is likely that further attempts to discover divergent geminiviruses in uncultivated hosts will prove successful.

The potential risks to cultivated crops of viruses that predominantly infect only uncultivated plant species has been documented for maize streak virus (MSV) (Varsani et al., 2008), the African Mastrevirus that causes MSD. MSV is the most economically important virus of maize in Africa. Maize was introduced to West Africa by the Portuguese in the early 1500s and to southern Africa by the Dutch in the mid-1600s but probably only began manifesting evidence of severe MSD around the 1860s (Monjane et al., 2011) with the emergence of a maize-adapted recombinant of two *Digitaria*-adapted MSV strains (Varsani et al., 2008). Therefore, besides purely taxonomic reasons for characterising geminiviruses that mainly infect uncultivated species, the ever present risk that such viruses can become adapted to and cause disease in cultivated hosts is a strong incentive for cataloguing the entire range of plant viral species that are found within terrestrial ecosystems.

Here we describe a new highly divergent geminivirus species isolated from the uncultivated South African spurge, *Euphorbia caput-medusae*. This new geminivirus has a unique genome organisation and distant sequence relatedness to other known geminiviruses and likely represents a new genus-level geminivirus lineage. Using an infectious genomic clone, we show that although it causes an asymptomatic infection in its uncultivated natural host, it can cause a severe infection in an important cultivated species such as tomato. The virus was named Euphorbia caput-medusae latent virus (EcmLV) and, accordingly, we propose that the new genus within which it should be placed be named Capulavirus. Together with a selection of diverse geminivirus sequences we use this new sequence to infer, firstly, previously undetected instances of likely inter-genus recombination in the geminiviruses and, secondly, that the most recent common ancestor of the geminiviruses was possibly a dicot-infecting virus with a *rep* gene that was expressed from a spliced complementary strand transcript.

2. Materials and methods

2.1. Plant sampling

In 2010, samples were collected in the Darling region of the Western Cape from *Euphorbia caput-medusae* plants as part of a large scale survey (for which >800 plants were collected) focusing on viral diversity at the interface between a preserved Cape fynbos ecosystem (Buffelsfontein Game and Nature Reserve) and an intensively cropped agro-ecosystem. Preliminary analysis of the collection of plants that were sampled showed that an unknown geminivirus was detected in an *E. caput-medusae* sample (see below). Nine more *E. caput-medusae* plants were collected in 2011 in the Western Cape region: three from the 2010 sampling site (Buffelsfontein Reserve), and six from coastal fynbos areas, including three near Laaiplek and three in Pater Noster (Supplementary Fig. 1 and Table 1). Whereas samples from the 2010 collection were preserved on dry ice before storage at −80 °C, those from 2011 were preserved by drying them with calcium chloride. The botanical identification of the 2010 plants was carried out initially by eye and was later confirmed by sequencing the C-terminal chloroplast *ndh*F gene using the primer pair 972-F (5′-GTC TCA ATT GGG TTA TAT GAT G-3′) and 2110-R (5′-CCC CCT AYA TAT TTG ATA CCT TCT CC-3′) (Kim and Jansen, 1995).

Table 1
Description of *Euphorbia caput-medusae* samples collected in 2010 and 2011 in the Western Cap floristic region of South Africa.

Sample name	Location	Sampling date	GPS position	PCR detection	Accession number
Dar10	Darling	Sept. 2010	33°14′45.96″S18°13′48.24″E	+	HF921459
Pan1	Pater Noster	Aug. 2011	32°48′25.81″S17°53′43.22″E	–	
Pan2	Pater Noster	Aug. 2011	32°48′26.17″S17°53′43.77″E	–	
Pan3	Pater Noster	Aug. 2011	32°48′28.12″S17°53′46.09″E	–	
Lap0	Laiiplek	Aug. 2011	32°42′36.17″S18°12′35.17″E	–	
Lap11	Laiiplek	Aug. 2011	32°42′36.36″S18°12′34.62″E	+	HF921477
Lap2	Laiiplek	Aug. 2011	32°42′37.06″S18°12′33.84″E	–	
Dar11	Darling	Aug. 2011	33°15′55.91″S18°12′58.63″E	+	HF921460
Dar0	Darling	Aug. 2011	33°15′55.05″S 18°13′0.47″E	–	
Dar2	Darling	Aug. 2011	33°15′56.19″S18°12′58.32″E	–	

2.2. DNA extraction, amplification, cloning and sequencing

Plant samples were ground with ceramic beads (MP 83 biomedicals) and purified quartz (Merck) within a grinding machine (Fastprep 24, MP biomedicals). Total DNA was extracted with the DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. Circular DNA molecules were amplified by RCA using *Phi29* DNA polymerase (TempliPhi™, GE Healthcare, USA) as previously described (Shepherd et al., 2008). The RCA product was digested with *Eco*RI, *Xho*I or *Bam*HI for 3 h at 37 °C; *Eco*RI and *Bam*HI generated a product of about 2.7 kbp. *Eco*RI restricted products of ~2.7 kbp obtained from two samples collected in Darling – one in 2010 (Dar10) and one in 2011 (Dar11) – were gel purified with the QIAquick Gel Extraction Kit (Qiagen), cloned into the pBC plasmid and sequenced. Sequence data were obtained by standard Sanger sequencing (Beckman Coulter Genomics) using a primer walking approach. Partially overlapping PCR primers were designed according to the sequences of clones Dar10 and Dar11 (Dar-1981F forward primer 5′-CCT CAC TGA ATC CAC ATC CA-3′ and Dar-1966R reverse primer 5′-CGA GGA ATT CGG ACT TGG-3′) to generate a third clone from a sample collected near Laaiplek in 2011 (Lap11), as follows: 1 μl RCA product obtained with Lap11 was amplified in a final volume of 25 μl containing 12.5 μl of HotStarTaq Plus Mastermix, 0.5 μl of each primer (at 10 μM concentration each) and 10.5 μl of RNase free water. The following amplification conditions were used: an initial denaturation at 95 °C for 5 min, followed by 30 cycles at 94 °C for 1 min, 60 °C for 1 min, 68 °C for 3 min, and a final extension step at 72 °C for 10 min. An amplification product of ~2.7 kbp was gel purified, ligated into pGEM-T Easy (Promega Biotech) and sequenced by standard Sanger sequencing using a primer walking approach.

2.3. Sequence analysis

Sequences were assembled using BioNumerics Applied Maths V6.5 (Applied. Maths, Ghent, Belgium) and compared to database sequences using BlastN, BlastP and tBlastX (Altschul et al., 1990). Open reading frames (ORFs) were identified that could potentially express proteins larger than 50 amino acids in length. Blast results were considered as indicative of significant homology when BLAST *e*-values were smaller than 10^{-2}.

The computer programme SMART (http://smart.embl-heidelberg.de/ (Letunic et al., 2012)) was also used for the detection of known domain architectures within the ORFs that had no detectable homologues within the public sequence databases. For nucleotide sequences and ORFs that did have clear homologues within other geminivirus genomes (those corresponding to the replication origins, transcription start/termination sites, replication associated protein and coat protein genes) we inferred the locations of domains and motifs within the newly sequenced genomes and their potentially expressed proteins based on the experimentally inferred positions of these domains in these other geminiviruses. Pairwise identity scores were calculated as previously described (Muhire et al., 2013).

2.4. PCR-mediated detection of EcmLV

Two 100% Dar10-complementary primer pairs were designed with PRIMER3 (Rozen and Skaletsky, 2000) which theoretically should not hybridise to 63 representative species of the family *Geminiviridae*. The first primer pair was designed to amplify a region starting within the large intergenic region and ending within the coat protein gene: Dar-136F forward primer 5′-CGA AGA GGT CAT TGG GAC AT-3′ and Dar-730R reverse primer 5′-CGG GTC TGG CTA AGA GAG TG-3′. The second primer pair is targeted to the *rep* gene: Dar-1662F forward primer 5′-TCG-ARC-AGG-TTT-CTG-GTC-CT and Dar-2257R reverse primer 3′-ACA-CCT-TCA-CTG-CCT-TGT-CC. The 9 *E. caput-medusae* samples collected in 2011 were PCR-tested with these two pairs of primers using the HotStarTaq Plus Master Mix Kit (Qiagen) following the manufacturer's protocol. Amplification conditions consist of an initial denaturation at 95 °C for 5 min, 30 cycles at 94 °C for 1 min, 55 °C for 1 min, 72 °C for 30 sec, and a final extension step at 72 °C for 10 min. A 100% pCAMBIA2300-complementary primer pair was designed with DNAMAN (version 5.0; Lynnon BioSoft, Quebec, Canada) in its 35S promoter region: pCambia2300F forward primer 5′-TGC TTT GAA GAC GTG GTT GG-3′ and pCambia2300R reverse primer 5′-ACG ACA CTC TCG TCT ACT CC-3′. Amplification conditions were as described above but with an annealing temperature of 65 °C.

2.5. Construction of an agro-infectious clone

To test the infectivity of the cloned geminivirus genome, we used a modification of the *Agrobacterium tumefaciens* delivery system (Peterschmitt et al., 1996) as follows: the insert of clone Dar10 was released using *Eco*RI and gel purified using a Wizard SV Gel and PCR Clean-Up System (Promega) following the manufacturer's protocol. A tandem head to tail dimer of the Dar10 genome was ligated into the dephosphorilated *Eco*RI cloning site of the binary vector pCAMBIA2300. This construct was introduced into *A. tumefaciens* C58 by electroporation.

2.6. Plant inoculation

A liquid culture of the Dar10 containing *A. tumefaciens* with an OD between 2.0 and 5.0 was concentrated ten times in sterile LB medium before inoculation. Test plants were inoculated either by injection with a syringe and a needle or by infiltration with the syringe alone directly in contact with the leaf. Alternatively a 24 h-plated culture of the Dar10 containing *A. tumefaciens* was used for inoculation as follows: the tip of an 18 Gauge ×1 1/2 needle previously dipped into the solid culture of the Dar10 containing *A. tumefaciens* was pricked several times into the plant. Ten tomato plants of the cv. Monalbo were inoculated by injection, ten by

infiltration and ten by pricking. A total of 30 *N. benthamiana* plants were similarly inoculated. Two or three plants of each species were not inoculated and used as non-infected controls. *E. caput-medusae* test plants were germinated from seeds by a commercial nursery. Due to the limited number of plants which could be supplied by the nursery, we tested different inoculation techniques on the same plants. Thus, each test plant was inoculated by both injection and pricking but on separate branches. Depending on the size of the plants, inoculation was done on one branch per technique (one plant), two branches per technique (one plant) or three branches per technique (three plants). Two *E. caput-medusae* plants were used as negative controls, one was inoculated with *A. tumefaciens* containing an empty vector and one was not inoculated. Plants were maintained in insect-free containment growth chambers under 14 h light at 26 °C, and 10 h dark at 24 °C.

2.7. Detection of potential inter-genus recombination events

We attempted to detect evidence of recombination within the Dar10 sequence by analyzing it together with nineteen other geminivirus full genome sequences collectively representing the most divergent geminivirus lineages available (see the "recombination analysis.rdp" and "recombination analysis alignment.fas" files provided as supplementary material for the identities of these 19 other viruses). These sequences were aligned with MEGA (Tamura et al., 2011) and analysed for recombination with the computer programme RDP4.18 (Martin et al., 2010) using the "verify alignment consistency" procedure outlined in Varsani et al. (2006). We analysed the alignment with seven of the recombination detection methods implemented in RDP4.18 with window size settings as follows: RDP = 100; BOOTSCAN = 800; MAXCHI = 240, CHIMAERA = 160; SISCAN = 500; 3SEQ and GENECONV with unspecified window sizes. These window sizes were set to approximately four times their default values to specifically detect, with a minimum of noise, signals of large sequence transfers (>300 nucleotides) between distantly related genomes. We discounted all signals of recombination that could not be confirmed (i) with four or more of the seven recombination detection methods and (ii) following realignment in isolation from the remainder of the 20 sequences in the alignment, of the sequence triplets within which signals were discovered. Whereas simulation analyses have revealed that the consensus of four methods should yield a false recombination detection rate per dataset of far below the 5% expected for each of the methods individually (Posada and Crandall, 2001), the realignment and retesting step ensured that the detected signals were not simply caused by alignment errors (Varsani et al., 2006).

2.8. Phylogenetic analysis

For purposes of reconstructing the evolutionary relationships of the various major geminivirus lineages we focused exclusively on the *rep* and *cp* coding regions of these genomes (i.e. the homologous portions). We assembled datasets consisting of 63 Rep and 59 CP amino acid sequences representing the entire breadth of known geminivirus diversity. Besides the inferred CP and Rep amino acid sequences of Dar10, each of these datasets contained 24 Mastrevirus sequences, 18 Begomovirus sequences, 8 Curtovirus sequences, 2 Becurtovirus sequences, one Eragrovirus sequence, one Turncurtovirus sequence, one Topocuvirus sequence, and one sequence each from divergent geminiviruses recently discovered infecting citrus plants (Citrus chlorotic dwarf associated virus, CCDaV, genome accession: JQ920490) (Loconsole et al., 2012) and grapevines (Grapevine Cabernet Franc-associated virus, GCFaV, genome accession: JQ901105) (Krenz et al., 2012). In addition to these sequences, the Rep dataset contained two geminivirus-like Rep sequences recently discovered within the apple genome (genome accession: PRJNA28845; whole genome shotguns: ACYM01134023 and ACYM01134026; Martin et al., 2011). Finally, the Rep dataset also contained two divergent outlier sequences that were used to root the Rep phylogeny: one derived from the witches broom associated phytoplasmal plasmid and the other from the geminivirus-like mycovirus Sclerotinia sclerotiorum hypovirulence-associated DNA virus (SsHADV, genome accession: NC_013116).

Because of the extremely distant relationships that existed even between the inferred amino acid sequences of these proteins it was not possible to accurately align the sequences. We therefore accounted for alignment uncertainty using a Bayesian approach to simultaneously estimate phylogenetic trees and alignments with the computer programme BAli-Phy version 2.2.1 (Suchard and Redelings, 2006). Default Bali-Phy settings were used for both the Rep and CP amino acid sequences. The LG2008 + gwF (Le and Gascuel, 2008) + generalised weighted frequencies amino acid substitution model (previously determined to be the best fit amino acid substitution model for these sequences (Varsani et al., 2009)) was used in conjunction with the RS07 (Suchard and Redelings, 2006) insertion-deletion model. Convergence was checked by comparing two independent MCMC chains for each of the two sets of aligned sequences. Estimated sample sizes (ESS) of model parameters were calculated using the programme Tracer http://tree.bio.ed.ac.uk/software/tracer, Tracerv 1.4 Rambaud and Drummond (2007) and ESS scores following merging of samples from the two independent chains run for each of the alignments were all greater than 200.

3. Results and discussion

3.1. Discovery of a new highly divergent geminivirus from E. caput-medusae

DNA was extracted from one *E. caput-medusae* plant, collected in Darling (Western Cape region, South Africa) in 2010. *E. caput-medusae* is a sprawling, succulent wild spurge indigenous to the coastal regions of South Western Africa. A 2.7 kbp *Eco*RI restricted DNA fragment was obtained from the RCA product generated from this plant. The *Eco*RI restricted fragment was cloned (clone Dar10), sequenced and shown to consist of 2678 bp. BlastN and BlastX comparisons between Dar10 and all sequences in Genbank indicated that the highest identity score was detected with a recently deposited geminivirus genome isolated from French bean in India: French bean severe leaf curl virus (FbSLCV; accession number NC_018453, highest percent identity = 78%, *e*-value = 1×10^{-169}). Although FbSLCV has not been described in the peer reviewed literature, a manuscript dealing with the biological characterisation of this virus is presently being prepared by its discoverers. Nevertheless, because the FbSLCV sequence is in the public domain, we included it in our comparative analyses between the Dar10-like viruses and the rest of the known geminiviruses. It is also noteworthy that in our scan of Genbank for sequences homologous with those occurring in Dar10, the second best match to Dar10 we found was to a geminivirus-like Rep sequence that is apparently integrated into the *Malus domestica* genome (Martin et al., 2011).

Among the nine *E. caput-medusae* plants collected in 2011, two tested positive by PCR for Dar10-like viruses: one from Darling and one from Laaiplek (Table 1). One viral DNA fragment was cloned from each of these two PCR-positive plants: clone Dar11 with 2650 bp and clone Lap11 with 2677 bp. The *Eco*RI cloning sites used for clones Dar10 and Dar11 was found to be unique in the Lap11 genome generated with partially overlapping primers. The *Eco*RI site was also confirmed to be unique in Dar10 and Dar11 with the sequenced PCR products generated with one of the

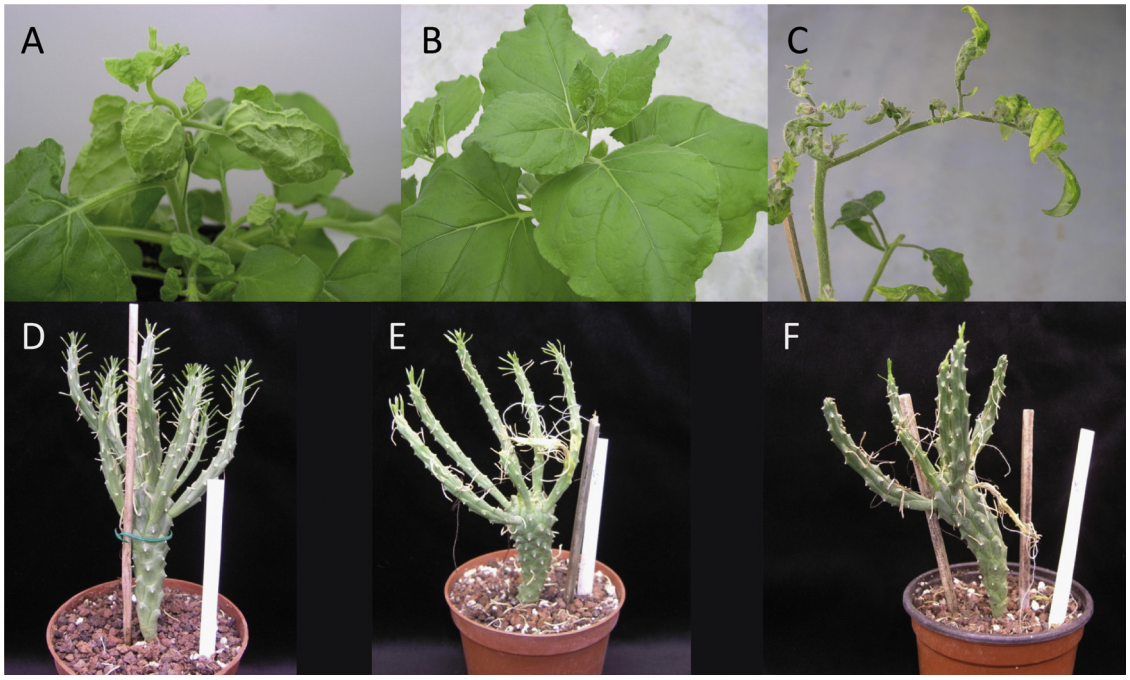


Fig. 1. Symptoms caused by EcmLV on *E. caput-medusae*, *N. benthamiana* and *S. lycopersicum* cv Monalbo plants at 21 dpi following agroinoculation with the agroinfectious clone Dar11. (A) *Nicotiana benthamiana* exhibited leaf curling, distortion and vein thickening. (B) *Nicotiana benthamiana*, which was used as non-infected control. (C) Tomato plants exhibited leaf curling, distortion, stunting and yellowing. (D) *Euphorbia caput-medusae* seedling which was used as non-infected control. (E) *E. caput-medusae* seedling agro-inoculated with the empty vector pCambia2300. It exhibited both yellowing wilt and necrosis on the branch inoculated with liquid culture. (F) *E. caput-medusae* seedling agro-inoculated with EcmLV and detected PCR-positive with EcmLV specific primers. The EcmLV inoculated plant did not exhibit any symptoms which could be related to viral infection.

specific primer pairs (Dar-1662F/Dar-2257R) spanning the *Eco*RI site. This indicated that the three cloned viral DNA fragments correspond to full length geminiviral genomes amplified by RCA. These circular DNA molecules were considered to be the complete viral genomes of geminiviruses infecting the three *E. caput-medusae* plants, because only one band was resolved by electrophoresis of the *Eco*RI and *Bam*HI digested RCA products (unique restriction sites for the viral clones) in the size range of geminivirus genome component sequences and no fragments were detected within the size ranges of known geminivirus satellite sequences. Similarly, restriction with *Xho*I, a non-cutting enzyme for any of the three clones, also did not show any DNA fragment in the satellite sequence size range.

To further confirm that the unique viral DNA sequences that we had cloned were complete genomes, we tested the infectivity of Dar10 in its natural host. One (plant 2) and two (plants 2 and 3) out of five *E. caput-medusae* plants which were agro-inoculated with clone Dar10, tested positive at 35 and 154 days post-inoculation (dpi), respectively, with the Dar10 specific primers, indicating that the clone was biologically active in its natural host. This finding suggests that the cloned 2.7 kb DNA fragment represents the complete genome of this new geminivirus (Supplementary Fig. 2A). The Dar10 positive samples were negative with a PCR detection test targeted to the pCAMBIA2300 binary vector, confirming that the positive virus detection was not simply due to the detection of the agro-inoculated construct (Supplementary Fig. 2B). The virus-infected plant did not exhibit any symptoms, which could differentiate it from the control plant which was agro-inoculated with the empty pCAMBIA2300 vector (Fig. 1). However we noticed that all inoculated plants reacted severely both to the injection of the liquid culture of *A. tumefaciens*, with yellowing wilt and necrosis of the inoculated branch, and to pricking inoculation with slight chlorosis and yellowing near the inoculation spot.

The lack of symptoms in the experimentally agro-infected plants at 35 and 154 dpi was consistent with the observation that the three wild plants of *E. caput-medusae* in which the new geminivirus was detected did not exhibit any conspicuous symptoms such as chlorotic mosaic, streaking or leaf deformation that differentiated them from plants that tested negative for Dar10-like viruses. Therefore, given that both field and experimentally infected *E. caput-medusae* plants did not exhibit any conspicuous symptoms (Fig. 1), we conclude that this geminivirus is likely latent in its natural host species, *E. caput-medusae*. This absence of visual symptoms is in line with results obtained during recent wild plant virus biodiversity surveys, which showed that most of the viral sequences discovered came from asymptomatic, healthy-looking plant samples (Melcher et al., 2008; Muthukumar et al., 2009; Roossinck et al., 2010).

Although the three full genomes isolated from the wild spurges could be aligned with the genome of FbSLCV, attempts to accurately align them with a diverse representation of other known geminiviruses proved largely unsuccessful due both to the extremely low degrees of sequence similarity in the homologous *cp* and *rep* genes, and the fact that the remainders of the genomes of these viruses were, with the exception of particular highly conserved sequence motifs (see below) not detectably homologous with those of other geminiviruses. The three sequences (Dar10; Dar11 and Lap10) showed between 93.6 and 95.3% pairwise identity scores. These identity scores are well above all of the species demarcation thresholds recommended for the various geminivirus genera by the geminivirus Study Group of the ICTV (Fauquet et al., 2008) and a recent proposal for the classification of viruses in the genus Mastrevirus (Muhire et al., 2013). The inter-clone distances are, however, unexpectedly high with respect to the relatively small geographical and temporal distances separating the three sampling sites (<60 km and approximately one year). The occurrence of long term infections in vegetatively propagated *E. caput-medusae* plants

and inefficient plant to plant viral transmission leading to deep population subdivisions, could both contribute to the presence of such divergent lineages within the sampling region. To test this hypothesis would however require additional virus samples and controlled infection and transmission experiments.

Although the geminivirus genomes isolated from *E. caput-medusae* plants are most closely related to FbSLCV, the pairwise identity score between them (71.7–72.5%) is below the lowest of the geminivirus species demarcation thresholds recommended by the ICTV (75% for mastreviruses) (Fauquet et al., 2008; Muhire et al., 2013). The name Euphorbia caput-medusae latent virus (EcmLV) is proposed for the new species.

3.2. Experimental hosts of EcmLV

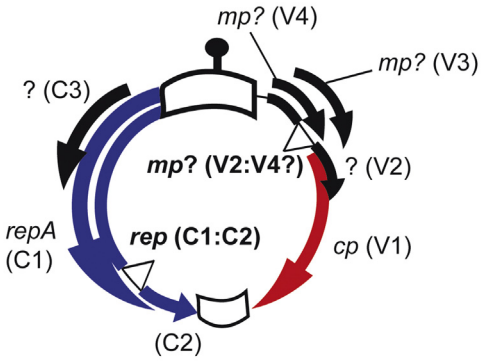
To test if EcmLV has the potential to infect species outside the family *Euphorbiaceae*, the Dar10 agroinfectious clone was inoculated into plants belonging to two solanaceous species, *Solanum lycopersicum* (the host of a large range of begomoviruses from the Old and New Worlds), and *Nicotiana benthamiana*, a very permissive host often used in experimental virology. Surprisingly, all 30 of the inoculated tomato plants exhibited severe symptoms 21 days post-inoculation, whatever the inoculation technique (injection, infiltration or pricking). The symptoms, similar to, but clearly distinct from those caused by tomato yellow leaf curl virus, consisted of curling and distortion of leaflets, leaf stunting, and prominent yellowing along leaf margins and/or interveinal regions (Fig. 1). Similarly, 27 out of 30 inoculated *N. benthamiana* plants (90%) exhibited severe symptoms 21 days post inoculation whatever the inoculation technique. Symptoms consisted of leaf curling and distortion and thickening of the veins (Fig. 1). Infected plants of both species were stunted and did not produce flowers. Whereas EcmLV was detected in symptomatic plants using the specific primers designed in this study, the asymptomatic plants and the non-inoculated plants tested negative for EcmLV.

EcmLV therefore potentially has a broad host range. This is consistent with the classical hypothesis that viruses evolving in species-rich-wild plant communities will likely adapt to infect a wide range of hosts – possibly even including those belonging to different plant families (Jones, 2009). There are however multiple factors which constrain virus infection in natural conditions, as has been revealed by a study in which five generalist virus species were each detected in a narrow range of host plants among twenty one wild species (Malpica et al., 2006).

3.3. Characteristics of the EcmLV genomes

The arrangement of open reading frames (ORF) (Fig. 2) within the 2678 bp circular DNA of the EcmLV-Dar10 clone is most similar to those reported for mastreviruses (Rosario et al., 2012), with two overlapping complementary sense ORFs (C1 and C2), and two intergenic regions: a large one (LIR) and a small one (SIR). A third ORF (C3) in the complementary sense region was also detected. Unlike mastrevirus genomes, which contain two virion sense ORFs (V1 and V2), the EcmLV circular DNA has four ORFs (V1, V2, V3 and V4) (Fig. 2) that would be predicted to encode proteins with >68 amino acids. The two other genomes obtained from Dar11 and Lap11 contained the same patterns of ORFs and intergenic regions. Except for the V4 ORF which was not detected in FbSLCV, the ORF pattern of FbSLCV is similar to that of EcmLV.

The V1 ORFs of the three EcmLV clones encode predicted proteins of 242 aa in length that share ~73% identity with the predicted coat protein of FbSLCV (*e*-value = 6×10^{-125}). The V2 ORFs of the three clones encode predicted proteins of 135 aa in length that share ~53% identity with the cognate ORF of FbSLCV. The V3 ORFs of the three clones encode proteins of 87 aa in length but the identity



ORF	coordinates	number of amino acids
V1 (*cp*)	546-1274	242
V2 (*mp*?)	129-335	68
V3	248-652	105
V4 (*mp*?)	169-432	87
C1 (*rep*A)	complement (1667-2431)	331
C2	complement (1297-1689)	130
C3	complement (1816-2181)	121

Fig. 2. Genomic organisation of EcmLV showing the arrangement of seven predicted open reading frames, putative proteins and the position of the small and large intergenic regions. Arrows indicate the positions and orientations of ORFs (V = virion sense and C = complementary sense) that are suspected to encode expressed proteins. *mp* = movement protein gene, *cp* = coat protein gene, *repA* = replication associated protein gene A, *rep*(C1:C2) = gene derived from a spliced transcript encompassing C1 and C2 ORFs, *mp?*(V2:V4) = gene derived from a spliced transcript encompassing V2 and V4 ORFs. A question mark indicates that an ORFs function is either completely unknown or only suspected. The only genes shared between all *Geminiviridae* genomes are *rep* (in blue) and *cp* (in red). Intergenic regions are represented as open blocks and the hairpin structure at the origin of virion strand replication is indicated at the 12 o'clock position. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

with the cognate ORF of FbSLCV is very low (36%). Except for their homologues encoded by FbSLCV, the V2, V3 and V4 ORFs of Dar10 potentially encode proteins with no significant degrees of identity with any other known protein.

Despite the absence of any detectable homologues of these proteins in sequence databases, exploration of protein domains and domain architectures using SMART (http://smart.embl-heidelberg.de/ (Letunic et al., 2012)) indicated that V3 and V4 contain 22 and 23 amino acid regions, respectively, that are likely transmembrane domains. Given that similar probable transmembrane domains are present in the movement protein encoded by mastreviruses (Boulton, 2002), this suggests that EcmLV V3 and V4 may encode movement proteins that are functional analogues (if not *bona fide* homologues) of those found in mastreviruses.

Similarly to the mastrevirus MSV in which an intron was detected in V2 (Wright et al., 1997), potential splice junctions were detected in the virion sense transcript of EcmLV. However, whereas the virion sense transcript results in an in-frame deletion (76 nucleotides for MSV (Wright et al., 1997)) of the movement protein gene of MSV, splicing of the EcmLV transcript would potentially result in both the elimination of the V2 ORF start codon and a frame shift that together would cause the V2 and V4 genes to be expressed as a single protein (Supplementary Fig. 3) in much the same way as full length *rep* genes are expressed from the two complementary sense ORFs of mastreviruses and becurtoviruses. Interestingly, similar likely splice junctions are present in the virion sense transcripts of the FbSLCV (Fig. 2). While suggesting that the involvement of virion sense gene transcript splicing as a gene expression strategy possibly predated the most recent common ancestor of EcmLV, FbSLCV, the mastreviruses and the begomoviruses, the specific

characteristics of this splicing in EcmLV would differentiate its genome architectures from those of all other known geminiviruses. It is noteworthy, however, that there is no homologue of the V4 ORF in the FbSLCV genome and, although potential virion strand transcript splice junctions are present in this virus, it is unlikely that a protein homologous to a V2–V4 protein of EcmLV could be expressed by this virus from spliced virion strand transcripts.

As is the case in mastreviruses and becurtoviruses, the EcmLV and FbSLCV genomes likely express a Rep protein from a spliced complementary sense transcript (Dekker et al., 1991); Fig. 2. The spliced transcript originating from EcmLV C1–C2 encodes a putative 331 aa protein which Exhibits 79% identity to the Rep protein of FbSLCV (*e*-value = 1×10^{-117}). It is also plausible that, as is likely the case in mastreviruses, the viruses also express a RepA protein of 254 aa from an unspliced complementary sense transcript. The inferred amino acid sequences of these EcmLV and FbSLCV Rep proteins contain canonical rolling circle replication (RCR) motifs which, besides being present in all known geminivirus Reps, are also strongly conserved amongst many other rolling circle replicons (Ilyina and Koonin, 1992; Koonin and Ilyina, 1993; Rosario et al., 2012) (Supplementary Fig. 4). In addition, the inferred EcmLV and FbSLCV Reps contain the ATPase motifs Walker-A, Walker-B, and C which were shown for other geminiviruses to be included in the Rep region exhibiting helicase activity (Choudhury et al., 2006; Clerot and Bernardi, 2006) (Supplementary Figs. 3 and 4). Unlike in some mastrevirus Reps and the Rep of the divergent geminivirus ECSV (new Eragrovirus genus), the predicted EcmLV and FbSLCV Rep proteins lack the canonical LXCXE retinoblastoma binding domain (Arguello-Astorga et al., 2004). The EcmLV and FbSLCV Reps do, however, contain the so-called GRS domain identified in other geminivirus Reps (Nash et al., 2011).

As in all other known geminiviruses, the 348–375 bp LIRs of the EcmLV isolates contain a predicted hairpin-loop structure with the extremely conserved geminiviral virion strand origin of replication nonanucleotide motif (TAATATTAC) in the loop (Lazarowitz et al., 1992) (Supplementary Fig. 3). Also similar to mastreviruses, the stem sequence of this predicted hairpin structure contains a number of nucleotide mismatches (Supplementary Fig. 5). Between the first probable TATA box of the complementary sense gene promoter (located at position 2540) and the replication origin, there is a GC-rich region that might, as is the case in other geminiviruses, constitute a G-Box with a role into transcriptional regulation (Eagle and Hanley-Bowdoin, 1997).

Between the likely hairpin structure and *rep* gene start codon are two different sets of directly and inversely repeated sequences arranged similarly to analogous "iteron" sequences that have been identified as *rep* binding recognition sites in begomoviruses, curtoviruses and topocuviruses (see the sequences labelled as Type A and Type B iterons in Supplementary Fig. 3) (Londono et al., 2010). The Lap11 LIR sequence differs somewhat from that of the Dar10 and Dar11 sequences in that it has 24 fewer nucleotides between the potential iteron sequence (type A) and the potential virion-sense gene TATA box.

Finally, the EcmLV and FbSLCV sequences respectively contain 21–23 bp and 93 bp small intergenic regions (SIR) between the *rep* and *cp* stop codons – a feature shared with mastreviruses, becurtoviruses, ECSV, CCDaV and GCFaV. In common with begomoviruses, topocuviruses and curtoviruses but unlike mastreviruses and ECSV, however, in EcmLV and FbSLCV the likely polyadenylation signals of the V-sense and C sense transcripts reside within the *rep* and *cp* genes.

3.4. Analysis of recombination

Since it has long been accepted that recombination between highly divergent geminiviruses could have played a role in the genesis of some of the known geminivirus genera (Briddon et al., 1996; Rybicki, 1994; Stanley et al., 1986) we attempted to identify evidence that EcmLV and/or FbSLCV might be inter-genus recombinants using a recombination–detection and alignment consistency verification approach previously devised by Varsani et al. (2006) for the analysis of recombination between extremely divergent sequences. Although this analysis yielded no evidence that either EcmLV or FbSLCV were recombinants, it both identified previously suspected recombination events within topocuvirus and curtovirus genomes (Briddon et al., 1996; Rybicki, 1994; Stanley et al., 1986; Varsani et al., 2009) and identified previously undetected evidence of recombination within the ECSV, CCDaV and Turnip curly top virus (TCTV) genomes (Fig. 3; see also the "recombination analysis.rdp" and "recombination analysis alignment.fas" files provided as supplementary material for additional details on these detected recombination events).

3.5. Phylogenetic relations between EcmLV and others geminiviruses

Possible evolutionary relationships between EcmLV and other known geminiviruses were investigated using phylogenetic analyses of inferred amino acid sequences from a representative sampling of geminiviral coat protein and Rep sequences. In the case of the Rep analysis, geminivirus-like phytoplasmal plasmid and mycovirus derived Rep amino sequences were included as outliers so that the phylogenetic tree determined with these amino acid sequences could be properly rooted. It was not possible to root the phylogenetic tree determined from CP amino acid sequences because there exist no suitable known non-geminiviral CP homologues. Also included in the Rep dataset was the geminivirus Rep-like sequence we had earlier found integrated within the apple genome.

Since the analysed amino acid sequences were so diverse, it was not possible to reliably align them. We therefore opted to use a Bayesian phylogenetic tree construction approach that explicitly accounts for alignment uncertainty by using a Markov chain Monte Carlo sampling scheme to simultaneously infer families of almost equally plausible alignments along with their associated phylogenetic trees. The maximum clade credibility (MCC) consensus of these trees is what we present in Fig. 4.

In both these trees, EcmLV clearly clusters with FbSLCV on a branch that is neither closely associated with any sequences classified within any of the seven established geminivirus genera, nor closely associated with any of the other known divergent geminivirus sequences. It is however, notable that the EcmLV and FbSLCV CP sequences are slightly more similar to those of the begomoviruses than they are to the other geminiviruses. However, given the lack of an outgroup, we cannot conclude that these viruses share a more recent common ancestor with the begomoviruses than they do with any of the other geminivirus groups.

The rooted Rep phylogeny on the other hand clearly indicated that EcmLV and FbSLCV cluster, with a posterior probability of 0.95, with all other geminiviruses that express Rep proteins from spliced complementary sense transcripts. Among the geminivirus Reps from known free-living geminiviruses (i.e. excluding Rep sequences integrated into host genomes) the EcmLV and FbSLCV Reps are most closely related to that from a recently discovered grapevine infecting geminivirus, GCFaV (Krenz et al., 2012).

It is also noteworthy that the sequences of apparently geminivirus-like Reps that are reportedly integrated into the apple genome (Martin et al., 2011) are very clearly most closely related to the EcmLV and FbSLCV sequences. While the possibility remains that these apparent integrons might be a sequence assembly artefacts arising due to the contamination of shotgun sequenced genomic apple DNA with DNA derived from an undetected

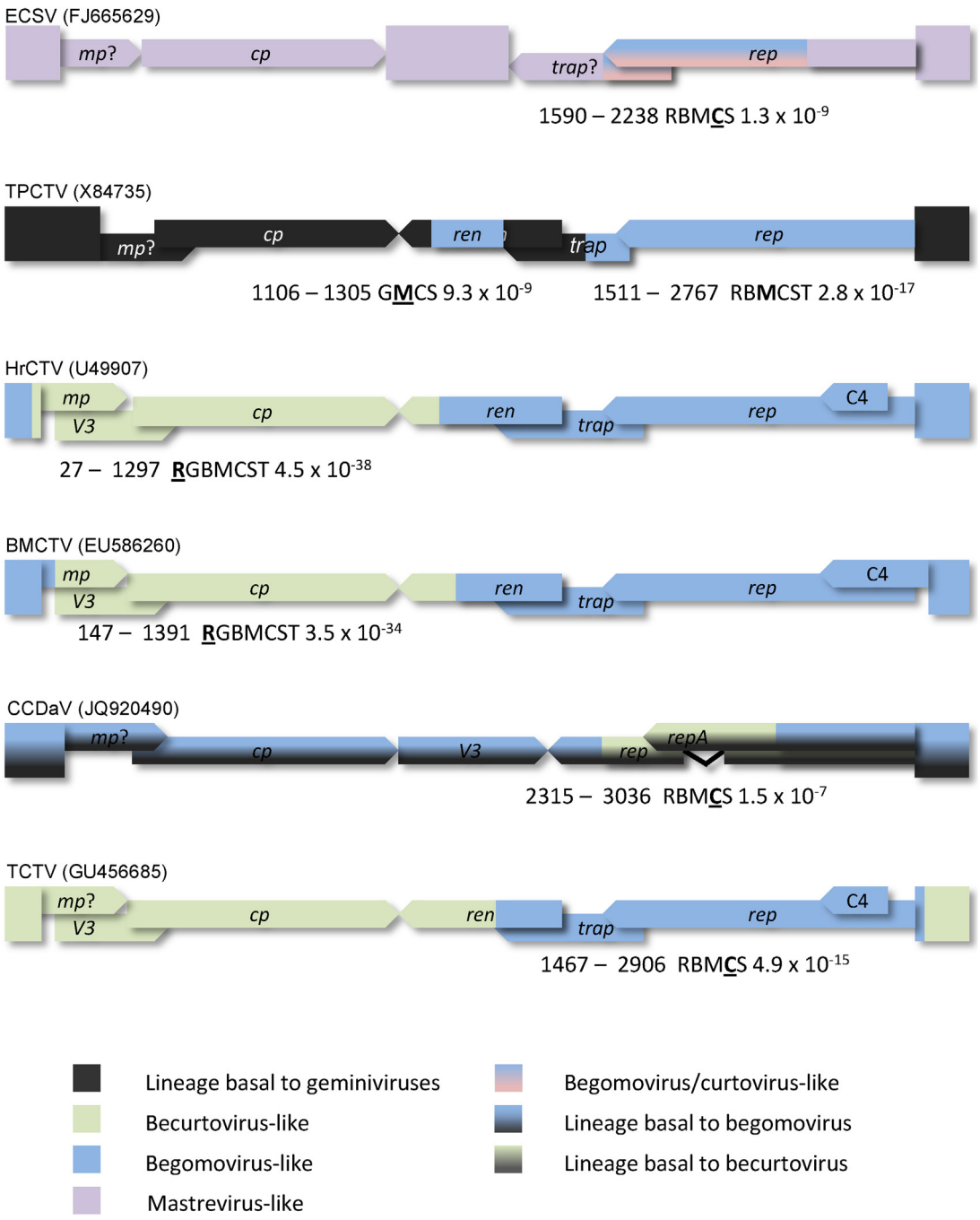


Fig. 3. Genome sequences of representative viruses with evidence of inter-genus recombination events. Bars indicate the genome sequences of representative inter-genus recombinants linearised at the virion strand origin of replication. Colours indicate the likely origins of the indicated genome regions. Also shown for each of the seven represented recombination events are the approximate beginning and ending coordinates of the recombinationally derived genome fragments (where nucleotide 1 is the first nucleotide 5′ of the virion strand origin of replication), the recombination analysis methods with which the recombination events are detectable (with an associated multiple testing corrected *p*-value <0.05; R = RDP, G = GENECONV, B = BOOTSCAN, M = MAXCHI, C = CHIMAERA, S = SISCAN, T = 3SEQ), and their associated *p*-value's (corresponding to the method indicated in bold/underlined). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

geminivral infection, it is nevertheless interesting that EcmLV-like viruses might, in addition to infecting members of the *Euphorbiaceae* and *Fabaceae*, also be capable of infecting members of the *Rosaceae*.

Finally it is important to point out that, even considering the unknown location of the CP tree root, the Rep and CP trees have major discrepancies that are potentially indicative of the Rep and CP sequences of many of the viruses considered here having separate evolutionary histories. It is apparent from previous studies and the analysis we describe above that recombination probably underlies the discordant locations of the various curtoviruses, ECSV and CCDaV in the two trees (Varsani et al., 2009). Crucially, the additional sequences that we included in our analysis relative to the study of Varsani et al. (2009), mean that we have for the first time been able to identify ECSV as having a genome that is mostly mastrevirus-like but with a fragment of the genome corresponding to the *rep* intron region of genuine mastreviruses having been derived from what appears to be a divergent begomovirus/curtovirus-like ancestor. While explaining the discordant position of ECSV in the Rep

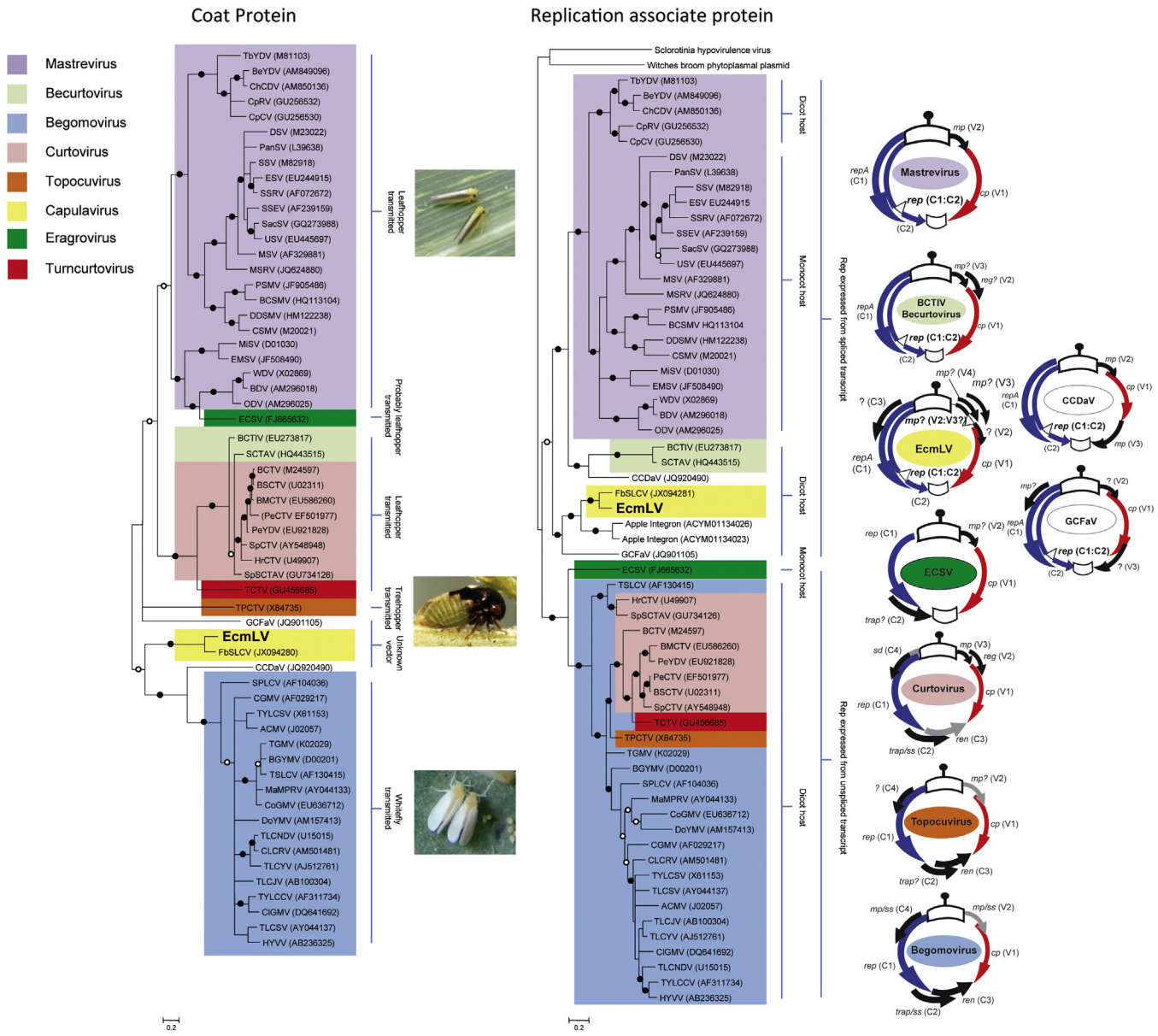


Fig. 4. Maximum clade credibility phylogenetic trees describing the evolutionary relationships between various geminivirus coat protein (CP) and replication associated protein (Rep) amino acid sequences. The trees were constructed so as to account for alignment uncertainty using simultaneous Bayesian inference of alignments and phylogenies. Species belonging to the seven established geminivirus genera (Mastrevirus, Becurtovirus, Topocuvirus, Begomovirus, Eragrovirus, Turncurtovirus and Curtovirus) and one tentative genus (Capulavirus) are indicated with coloured blocks. Branches with a filled dot have >99% posterior probability support whereas those with an empty dot have >95% posterior probability support. All branches with less than 80% posterior probability support have been collapsed. Indicated on the CP tree are the known/likely insect vectors of many of the viruses. Indicated on the Rep tree are the hosts of the viruses, the genomic organisation of the different geminivirus genera and divergent geminiviruses, and whether or not their Rep proteins are expressed from spliced complementary strand transcripts. Photo courtesy of: J.M. Lett from CIRAD (*Cicadulina mbila*), John Innes Centre (*Micrutalis malleifera* Fowler) and A. Franck from CIRAD (*Bemisia tabaci*).

and CP phylogenies this recombination event could also explain why, in common with begomoviruses and curtoviruses, it has a Rep that is expressed from an unspliced complementary strand transcript.

3.6. What does the new data tell us about the earliest geminiviruses?

EcmLV and the various highly divergent geminiviruses represented within our rooted Rep phylogeny permit us to infer, with the greatest accuracy yet achieved, some of the likely characteristics of the earliest geminiviruses. It is noteworthy, for example, that the monocot-infecting mastreviruses form a well-supported (with a 0.99 posterior probability) monophyletic clade within the Rep phylogeny that is nested within a much larger more divergent cluster of dicot-infecting geminiviruses with spliced Reps. It is therefore most likely that the most recent common ancestor (MRCA) of the mastreviruses infected dicots and that the host switch occurred from dicots to monocots and not the other way around as has been previously supposed (Varsani et al., 2009). The only known non-mastreviral monocot-infecting geminivirus is ECSV which, because it branches near the root of the geminivirus Rep phylogeny, might be interpreted as evidence that the MRCA of the geminiviruses could have plausibly infected either monocots or dicots (or perhaps even the common ancestor of these plant lineages). It is, however, apparent

that the basal location of ECSV in the Rep tree is potentially due to this isolate having a recombinant *rep* gene that is approximately 1/3 mastrevirus-like and 2/3 begomovirus/curtovirus-like (Fig. 3). If one accounts for this it would imply that the basal-position of ECSV in the Rep phylogeny may be largely artefactual and that the "true" position of the majority of its genome is at the base of the monocot-infecting mastrevirus lineage (as it is in the CP tree). If this is the case then one need invoke only a single dicot to monocot host switch to explain the present host distributions of all the known geminiviruses.

Given the clear separation within the Rep phylogeny of viruses with and without a *rep* intron (Fig. 4), it is similarly possible to infer that in geminiviruses there was possibly only a single instance of either mutational loss or gain of the *rep* gene intron splicing signals. It is, however, not entirely clear whether the MRCA of the geminiviruses had a *rep* gene intron or not since the viral lineages with both types of gene branch from the root of the *rep* gene phylogeny. It is perhaps significant that five of the six inter-genus recombinants identified by our recombination analyses (Fig. 3) have "intronless" *rep* genes and carry evidence of undergoing recombination events that "converted" a virus with a spliced *rep* gene into a virus with an unspliced *rep* gene. The exceptional case, CCDaV, appears to have involved the conversion of a virus with an intronless *rep* gene into one that had a *rep* gene intron. It is also noteworthy that some species of the geminivirus-like mycoviruses (represented in Fig. 4 by SsHADV) also likely express Rep proteins from spliced complementary strand transcripts (Dayaram et al., 2012) and it is conceivable therefore that the *rep* gene of the MRCA of the geminiviruses and geminivirus-like mycovirus could have also contained an intron.

3.7. *Capulavirus: a new genus of the Geminiviridae family*

EcmLV and FbSLCV clearly belong to a highly divergent geminivirus lineage. Moreover, the virion-sense genes of these new viruses exhibit a unique organisation among this family with only the *cp* gene having detectable homologues in other currently known geminivirus genomes. Given that these viruses are obviously distinguishable from the other established geminivirus genera based on sequence relatedness and genome organisation, we suggest that EcmLV and FbSLCV be placed within a new geminivirus genus. The name that we propose for this new genus is "Capulavirus". It is expected that Capulavirus may also be distinguishable based on the vector criteria, because the vector of EcmLV, if any, is expected to be different from the vectors so-far reported for other geminiviruses. Indeed, the coat proteins of EcmLV and FbSLCV, the only protein likely to be involved in their vector specificity (Briddon et al., 1990) is very distantly related to all other previously reported geminivirus coat proteins. Further studies are needed to confirm this expectation and additionally to test whether EcmLV and FbSLCV are also distinct from the other genera with respect to the breadth and/or specificity of the range of host species that they naturally infect.

Acknowledgements

This work was supported by Fondation pour la Recherche sur la Biodiversité, Direction Générale de l'Armement (Ministère de la Défense, France), Métaprogramme INRA « Meta-omics of microbial ecosystems » and CIRAD. We wish to express our sincere thanks and appreciation to Mr Paul Loubser and colleagues from Buffelsfontein Game & Nature Reserve.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2013.07.006.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215 (3), 403–410.
Arguello-Astorga, G., Lopez-Ochoa, L., Kong, L.J., Orozco, B.M., Settlage, S.B., Hanley-Bowdoin, L., 2004. A novel motif in geminivirus replication proteins interacts with the plant retinoblastoma-related protein. Journal of Virology 78 (9), 4817–4826.
Boulton, M.I., 2002. Functions and interactions of mastrevirus gene products. Physiological and Molecular Plant Pathology 60 (5), 243–255.
Briddon, R.W., Bedford, I.D., Tsai, J.H., Markham, P.G., 1996. Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology 219 (2), 387–394.
Briddon, R.W., Heydarnejad, J., Khosrowfar, F., Massumi, H., Martin, D.P., Varsani, A., 2010. Turnip curly top virus, a highly divergent geminivirus infecting turnip in Iran. Virus Research 152 (1/2), 169–175.
Briddon, R.W., Pinner, M.S., Stanley, J., Markham, P.G., 1990. Geminivirus coat protein gene replacement alters insect specificity. Virology 177 (1), 85–94.
Choudhury, N.R., Malik, P.S., Singh, D.K., Islam, M.N., Kaliappan, K., Mukherjee, S.K., 2006. The oligomeric Rep protein of Mungbean yellow mosaic India virus (MYMIV) is a likely replicative helicase. Nucleic Acids Research 34 (21), 6362–6377.
Clerot, D., Bernardi, F., 2006. DNA helicase activity is associated with the replication initiator protein rep of tomato yellow leaf curl geminivirus. Journal of Virology 80 (22), 11322–11330.
Dayaram, A., Opong, A., Jaschke, A., Hadfield, J., Baschiera, M., Dobson, R.C.J., Offei, S.K., Shepherd, D.N., Martin, D.P., Varsani, A., 2012. Molecular characterisation of a novel cassava associated circular ssDNA virus. Virus Research 166 (1/2), 130–135.
Dekker, E.L., Woolston, C.J., Xue, Y.B., Cox, B., Mullineaux, P.M., 1991. Transcript mapping reveals different expression strategies for the bicistronic Rnas of the geminivirus wheat dwarf virus. Nucleic Acids Research 19 (15), 4075–4081.
Delwart, E., 2012. Animal virus discovery: improving animal health, understanding zoonoses, and opportunities for vaccine development. Current Opinion in Virology 2 (3), 344–352.
Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. Journal of Virology 82 (2), 957–965.
Eagle, P.A., Hanley-Bowdoin, L., 1997. Cis elements that contribute to geminivirus transcriptional regulation and the efficiency of DNA replication. Journal of Virology 71 (9), 6947–6955.
Fauquet, C.M., Briddon, R.W., Brown, J.K., Moriones, E., Stanley, J., Zerbini, M., Zhou, X., 2008. Geminivirus strain demarcation and nomenclature. Archives of Virology 153 (4), 783–821.
Fauquet, C.M., Stanley, J., 2003. Geminivirus classification and nomenclature: progress and problems. Annals of Applied Biology 142 (2), 165–189.
Fuller, C., 1901. Mealie variegation. First report of the government entomologist 1899–1900., pp. 17–19.
Ge, L., Zhang, J., Zhou, X., Li, H., 2007. Genetic structure and population variability of Tomato yellow leaf curl China virus. Journal of Virology 81 (11), 5902–5907.
Haible, D., Kober, S., Jeske, H., 2006. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. Journal of Virological Methods 135 (1), 9–16.
Ilyina, T.V., Koonin, E.V., 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Research 20 (13), 3279–3285.
Isnard, M., Granier, M., Frutos, R., Reynaud, B., Peterschmitt, M., 1998. Quasispecies nature of three maize streak virus isolates obtained through different modes of selection from a population used to assess response to infection of maize cultivars. Journal of General Virology 79 (Pt 12), 3091–3099.
Jeske, H., 2009. Geminiviruses. In: zur Hausen, H., de Villiers, E.-M. (Eds.), Torque Teno Virus: The Still Elusive Human Pathogens, vol. 331. Springer, Berlin, pp. 185–226.
Jones, R.A.C., 2009. Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Research 141 (2), 113–130.
Kim, K.J., Jansen, R.K., 1995. Ndhf sequence evolution and the major clades in the sunflower family. Proceedings of the National Academy of Sciences of the United States of America 92 (22), 10379–10383.
Koonin, E.V., Ilyina, T.V., 1993. Computer-assisted dissection of rolling circle DNA replication. Biosystems 30 (1–3), 241–268.
Krenz, B., Thompson, J.R., Fuchs, M., Perry, K.L., 2012. Complete genome sequence of a new circular DNA virus from grapevine. Journal of Virology 86 (14), 7715.
Lazarowitz, S.G., Wu, L.C., Rogers, S.G., Elmer, J.S., 1992. Sequence-specific interaction with the viral AL1 protein identifies a geminivirus DNA replication origin. Plant Cell 4 (7), 799–809.
Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25 (7), 1307–1320.

Legg, J.P., Fauquet, C.M., 2004. Cassava mosaic geminiviruses in Africa. Plant Molecular Biology 56 (4), 585–599.
Letunic, I., Doerks, T., Bork, P., 2012. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Research 40 (D1), D302–D305.
Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G.P., Saponari, M., 2012. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family *Geminiviridae*. Virology 432 (1), 162–172.
Londono, A., Riego-Ruiz, L., Arguello-Astorga, G.R., 2010. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Archives of Virology 155 (7), 1033–1046.
Malpica, J.M., Sacristan, S., Fraile, A., Garcia-Arenal, F., 2006. Association and host selectivity in multi-host pathogens. PLoS ONE 1, e41.
Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011. Recombination in eukaryotic single stranded DNA viruses. Viruses 3 (9), 1699–1738.
Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26 (19), 2462–2463.
Martin, D.P., Shepherd, D.N., 2009. The epidemiology, economic impact and control of maize streak disease. Food Security 1 (3), 305–315.
Melcher, U., Muthukumar, V., Wiley, G.B., Min, B.E., Palmer, M.W., Verchot-Lubicz, J., Ali, A., Nelson, R.S., Roe, B.A., Thapa, V., Pierce, M.L., 2008. Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from *Ambrosia psilostachya*. Journal of Virological Methods 152 (1-2), 49–55.
Moffat, A.S., 1999. Plant pathology – Geminiviruses emerge as serious crop threat. Science 286 (5446), 1835-1835.
Monjane, A.L., Harkins, G.W., Martin, D.P., Lemey, P., Lefeuvre, P., Shepherd, D.N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E.K., Lakoutene, D.P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J.B., van Antwerpen, T., Owor, B.E., Flett, B., Ramusi, M., Windram, O.P., Syed, R., Lett, J.M., Briddon, R.W., Markham, P.G., Rybicki, E.P., Varsani, A., 2011. Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. Journal of Virology 85 (18), 9623–9636.
Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V.G., Briddon, R.W., Varsani, A., 2013. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family *Geminiviridae*). Archives of Virology 158 (6), 1411–1424.
Muthukumar, V., Melcher, U., Pierce, M., Wiley, G.B., Roe, B.A., Palmer, M.W., Thapa, V., Ali, A., Ding, T., 2009. Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. Virus Research 141 (2), 169–173.
Nash, T.E., Dallas, M.B., Reyes, M.I., Buhrman, G.K., Ascencio-Ibanez, J.T., Hanley-Bowdoin, L., 2011. Functional analysis of a novel motif conserved across geminivirus Rep proteins. Journal of Virology 85 (3), 1182–1192.
Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011a. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS ONE 6 (4).
Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y.J., Rohwer, F., Breitbart, M., 2011b. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS ONE 6 (6).
Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265 (2), 218–225.
Patil, B.L., Fauquet, C.M., 2009. Cassava mosaic geminiviruses: actual knowledge and perspectives. Molecular Plant Pathology 10 (5), 685–701.
Peterschmitt, M., Granier, M., Frutos, R., Reynaud, B., 1996. Infectivity and complete nucleotide sequence of the genome of a genetically distinct strain of maize streak virus from Reunion Island. Journal of Virology 141 (9), 1637–1650.
Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America 98 (24), 13757–13762.
Rambaud, A., Drummond, A.J., 2007. Tracer, version 1.4. http://beast.bio.ed.ac.uk/Tracer
Rey, M.E.C., Ndunguru, J., Berrie, L.C., Paximadis, M., Berry, S., Cossa, N., Nuaila, V.N., Mabasa, K.G., Abraham, N., Rybicki, E.P., Martin, D., Pietersen, G., Esterhuizen, L.L., 2012. Diversity of dicotyledenous-infecting geminiviruses and their associated DNA molecules in southern Africa, including the South-west Indian ocean islands. Viruses 4 (9), 1753–1791.
Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G.A., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Molecular Ecology 19, 81–88.
Rosario, K., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. Archives of Virology 157 (10), 1851–1871.
Rozen, S., Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology 132, 365–386.
Rybicki, E.P., 1994. A phylogenetic and evolutionary justification for three genera of *Geminiviridae*. Archives of Virology 139 (1/2), 49–77.
Rybicki, E.P., Pietersen, G., 1999. Plant virus disease problems in the developing world. Advances in Virus Research 53 (53), 127.
Schubert, J., Habekuss, A., Kazmaier, K., Jeske, H., 2007. Surveying cereal-infecting geminiviruses in Germany – diagnostics and direct sequencing using rolling circle amplification. Virus Research 127 (1), 61–70.
Shepherd, D.N., Martin, D.P., Lefeuvre, P., Monjane, A.L., Owor, B.E., Rybicki, E.P., Varsani, A., 2008. A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. Journal of Virological Methods 149 (1), 97–102.
Stanley, J., Markham, P.G., Callis, R.J., Pinner, M.S., 1986. The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. EMBO Journal 5 (8), 1761–1767.
Suchard, M.A., Redelings, B.D., 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22 (16), 2047–2048.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28 (10), 2731–2739.
Tan, P.H.N., Wong, S.M., Wu, M., Bedford, I.D., Saunders, K., Stanley, J., 1995. Genome organization of ageratum yellow vein virus, a monopartite whitefly-transmitted geminivirus isolated from a common weed. Journal of General Virology 76, 2915–2922.
Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P., Martin, D.P., 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. Virology Journal 6, 36.
Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Lefeuvre, P., Lett, J.M., Martin, D.P., 2008. Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. Journal of General Virology 89 (Pt. 9), 2063–2074.
Varsani, A., van der Walt, E., Heath, L., Rybicki, E.P., Williamson, A.L., Martin, D.P., 2006. Evidence of ancient papillomavirus recombination. Journal of General Virology 87, 2527–2531.
Warburg, O.M.D.S., 1894. Die kulturpfl anzen usambaras. Mitteilungen aus den DeutschenSchutzgebieten 7, 131.
Wright, E.A., Heckel, T., Groenendijk, J., Davies, J.W., Boulton, M.I., 1997. Splicing features in maize streak virus virion- and complementary-sense gene expression. Plant Journal 12 (6), 1285–1297.
Yazdi, H.R.B., Heydarnejad, J., Massumi, H., 2008. Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. Virus Genes 36 (3), 539–545.

2.2. Identification et caractérisation de nouveaux génomes de Capulavirus

Nos travaux de métagénomique réalisés en Camargue ont permis de mettre en évidence la présence de deux génomes viraux proches d'EcmLV sur des plants de luzerne cultivée (*Medicago sativa*), l'un en 2010 et l'autre en 2012. De plus, d'autres travaux de métagénomique effectués en Finlande par l'équipe d'Anna-Liisa Laine ont également permis de détecter la présence d'un génome viral proche d'EcmLV sur un échantillon de plantain (*Plantago lanceolata*). La caractérisation de ces génomes nous est alors parue essentielle pour mieux décrire le nouveau genre Capulavirus.

2.2.1. Matériel et méthodes

2.2.1.1. Plantes sources

- Plantain (*Plantago lanceolata* ; Finlande) : En 2013, 144 plants sauvages de *Plantago lanceolata* (*Plantaginaceae*) ont été collectés dans l'archipel d'Åland au Sud-Ouest de la Finlande. Une analyse du virome de ces plantes a été réalisée par une « approche siRNA » comme décrit par Candresse *et al.* (Candresse *et al.*, 2014). L'échantillon à partir duquel nous avons détecté la présence potentielle d'un capulavirus a été conservé à -80°C.

- Luzerne (*Medicago sativa* ; France) : Lors de nos travaux de géo-métagénomique en Camargue nous avons détecté la présence potentielle de capulavirus sur deux plants indépendants de luzernes: l'un collecté en 2010 (nommé 44.1E) et l'autre collecté en 2012 (nommé 48.2A). Ces deux échantillons ont été conservés à -80°C.

2.2.1.2. Extractions d'ADN, amplification, clonage et séquençage

Les ADN totaux de l'échantillon de plantain et des deux échantillons de luzerne ont été extraits selon la méthode décrite par Bernardo *et al.* (Bernardo *et al.*, 2013). Les ADN circulaires contenus dans les extraits de luzerne ont été amplifiés par RCA comme décrit par Shepherd *et al.* (Shepherd *et al.*, 2008). Pour tenter de linéariser les ADN circulaires ainsi amplifiés, ils ont été soumis à des digestions indépendantes par *Dra*I, *Nco*I et *Nde*I pour les échantillons issus de luzerne et par *Afl*II pour le plantain durant 3h à 37°C. Une approche alternative a été développée pour obtenir des génomes complets linéaires issus de RCA. Elle consiste à une amplification PCR à l'aide des amorces chevauchantes Dar-1981F et Dar-1966R, conçues à partir des contigs générés à partir de nos données de métagénomique (amorces et cycles PCR en Annexe 7). Les produits PCR ont été clonés dans pGEM-T Easy (Promega Biotech) selon la méthode décrite par Bernardo *et al.* (Bernardo *et al.*, 2013). Concernant l'échantillon de plantain, le « scaffold » issu de l'assemblage *de novo* des reads de métagénomique nous a permis de dessiner 8 paires d'amorces chevauchantes permettant de couvrir le génome viral total (conditions PCR et amorces en Annexe 7 ; les amorces en question sont précédées du suffixe Plantago). Les clones viraux issus des échantillons de luzerne et les produits PCR

issus de l'échantillon de plantain ont été séquencés via la méthode Sanger (Beckman Coulter Genomics).

2.2.1.3. Analyse de séquences

Les séquences ont été assemblées en utilisant le logiciel Bio Numerics Applied Maths V6.5 (Applied. Maths, Ghent, Belgium) et comparées au séquences de la GenBank par BlastN, BlastP et tBlastX (Altschul *et al.*, 1990). Les ORFs qui pouvaient potentiellement exprimer des protéines de plus de 50 acides aminés ont été détectés par le logiciel ORF finder de NCBI et identifiées via BlastP. Les résultats issus des Blasts ont été considérés comme significatifs lorsque la e-value était inférieure à 10^{-2}. La détection de domaines connus à l'intérieur des ORFs pour lesquelles nous n'avons détecté aucun homologue à partir de la collection GenBank a été effectuée comme décrit par Bernardo *et al.* (Bernardo *et al.*, 2013).

2.2.1.4. Démarcation d'espèces

Nous avons étudié la démarcation en espèces entre les différents isolats de capulavirus dont nous disposions à l'aide du logiciel SDT v1.0 (Muhire *et al.*, 2013). Au total nous disposions de 21 génomes de capulavirus dont 1 issu du plantain, 2 de la luzerne, 2 de FbSLCV, et 16 d'EcmLV ; pour l'obtention des génomes entiers des différents isolats d'EcmLV, se référer à la section 3.2.5.

2.2.1.5. Analyse de recombinaison

Une analyse de recombinaison entre les génomes entiers de capulavirus et 19 autres génomes de géminivirus représentatifs de la famille *Geminiviridae* a été entreprise selon la méthode décrite par Bernardo *et al.* (Bernardo *et al.*, 2013).

2.2.1.6. Analyse phylogénétique

Dans le but de reconstruire les relations évolutives entre les différentes lignées majeures de géminivirus, nous nous sommes focalisés sur les séquences peptidiques de la Rep et de la CP. Le jeu de données « CP » est constitué de 45 CP de géminivirus représentatifs de la famille *Geminiviridae* et de 5 CP de capulavirus (2 d'EcmLV, 1 de FbSLCV, 2 du virus de luzerne et 1 du virus du plantain). Le jeu de données, « Rep » est constitué de 64 Rep de géminivirus représentatifs de la famille *Geminiviridae* et de 5 Rep de capulavirus comme décrits pour la CP. Nos deux jeux de données ont alors été alignés via la méthode ClustalW (Thompson *et al.*, 1994) implémentée dans MEGA 5.2.1 (Tamura *et al.*, 2011). A partir de ces alignements, des analyses phylogénétiques par maximum de vraisemblance ont été réalisées grâce au logiciel PhyML 3.1 (Guindon *et al.*, 2005) selon le modèle évolutif WAG+I+G. L'arbre phylogénétique a été construit en réalisant 500 bootstraps.

2.2.2. Résultats

2.2.2.1. Découverte de trois nouveaux capulavirus

- ***Plantain***

Les travaux de siRNA-métagénomique réalisés en Finlande ont permis de détecter la présence d'un nouveau capulavirus sur plantain. L'extraction d'ADN suivie d'une RCA et d'une digestion via l'enzyme *Afl*II nous a permis d'obtenir un fragment unique de 2.8kpb laissant supposer que ce génome viral n'est pas accompagné d'un satellite (la taille des satellites de géminivirus est d'environ 1,3kb). Les PCR via les amorces chevauchantes nous ont permis d'obtenir le génome entier de ce virus qui fait 2832pb. Les comparaisons effectuées à partir de BlastN et de BlastX entre ce génome et les génomes archivés dans la GenBank ont montré que ce virus présente le pourcentage d'identité le plus élevé avec EcmLV (pourcentage d'identité=72%, E-value=1.10^{-60}). Le virus que représente ce nouveau génome n'a pas pu être nommé car pour le moment nous n'avons aucune information sur son possible effet symptomatique sur sa plante hôte. D'une part il est difficile de détecter des symptômes sur des morceaux de plantes congelées, et surtout, nous n'avons pas pu démontrer le postulat de Koch en l'absence d'un clone infectieux. C'est pourquoi nous l'appellerons provisoirement « *Plantago Capulavirus* ». Bien que le génome de *Plantago Capulavirus* soit très proche de celui d'EcmLV et de FbSLCV, leur score d'identité deux à deux (Figure 2.1, 59.4-61.8%) est en dessous du seuil de démarcation de l'espèce le plus bas indiqué par l'ICTV (75% pour les mastrevirus) (Fauquet *et al.*, 2008) et par Muhire *et al.* (Muhire *et al.*, 2013). *Plantago Capulavirus* est donc manifestement une nouvelle espèce de géminivirus.

- ***Luzerne***

La métagénomique nous a permis de détecter la présence potentielle de deux nouveaux capulavirus sur deux échantillons de luzernes indépendants collectés dans deux localités différentes, l'un en 2010 (nommé 44.1E) et l'autre en 2012 (nommé 48.2A). Ces échantillons ont été soumis à une RCA suivi de digestions enzymatiques indépendantes via *Dra*I, *Nco*I et *Nde*I. Chaque digestion enzymatique a permis d'obtenir des produits d'environ 2.8kpb. De plus nous avons séquencé les clones obtenus à partir de produits PCR réalisées à l'aide d'amorces chevauchantes. Ainsi nous avons pu observer que les sites de restriction *Dra*I, *Nco*I et *Nde*I étaient uniques au sein de ces génomes, ce qui nous a permis de confirmer que ces génomes n'étaient pas accompagnés de satellites. Nous avons ainsi obtenu des séquences de 2747pb pour le génome viral issu de 44.1E et 2766pb pour celui issu de 48.2A. Les comparaisons réalisées avec BlastN entre 44.1E, 48.2A et les séquences de la GenBank nous ont indiqué que la séquence la plus proche de ces génomes était celle d'EcmLV avec des pourcentages d'identité respectifs de 75% et 76% et des e-values respectives de $3x10^{-120}$ et $4x10^{-144}$. Les deux génomes viraux issus de 44.1E et 48-2A ont 82.2% d'identité deux à deux ce qui suggère qu'ils représentent des isolats appartenant à une même espèce de géminivirus. Toutefois, les scores d'identité deux à deux de ces deux génomes et de ceux de EcmlV, FbslcV et *Plantago Capulavirus* s'étendent de 59.9% à 70.2% (Figure 2.1). Ces

scores sont en dessous du seuil de démarcation de l'espèce le plus bas indiqué par l'ICTV (75% pour les mastrevirus) (Fauquet *et al.*, 2008) et par Muhire *et al.* (Muhire *et al.*, 2013). Les deux isolats viraux issus de la luzerne appartiendraient donc à une nouvelle espèce de géminivirus.

Pour essayer de déterminer si un symptôme particulier pouvait être associé à la présence de ce virus, nous avons réalisé en mars 2014une campagne d'échantillonnage de plantes de luzerne à l'aveugle, i.e. sans tenir compte de l'aspect des plantes (cf. section 3). Ces plantes ont été soumises à un test de détection via les amorces Luz-CP-F et Luz-CP-R. Les plantes « positives » par PCR étaient toutes chétives avec des feuilles incurvées et boursouflées, des symptômes qui n'étaient pas présents sur les plantes négatives. Pour confirmer l'association de ces symptômes à la détection positive par PCR, une deuxième campagne d'échantillonnage a été effectuée en août 2014 en ne prélevant que des plantes exhibant ce type de symptôme. Etant donné que 91% de ces plantes symptomatiques étaient positives par PCR pour la détection virale, .nous supposons que les symptômes d'enroulement, boursouflement et jaunissement des feuilles ainsi que la chétivité des plants sont caractéristiques de ce nouveau virus que nous proposons d'appeler *Alfafa leaf curl virus* (ALCV).

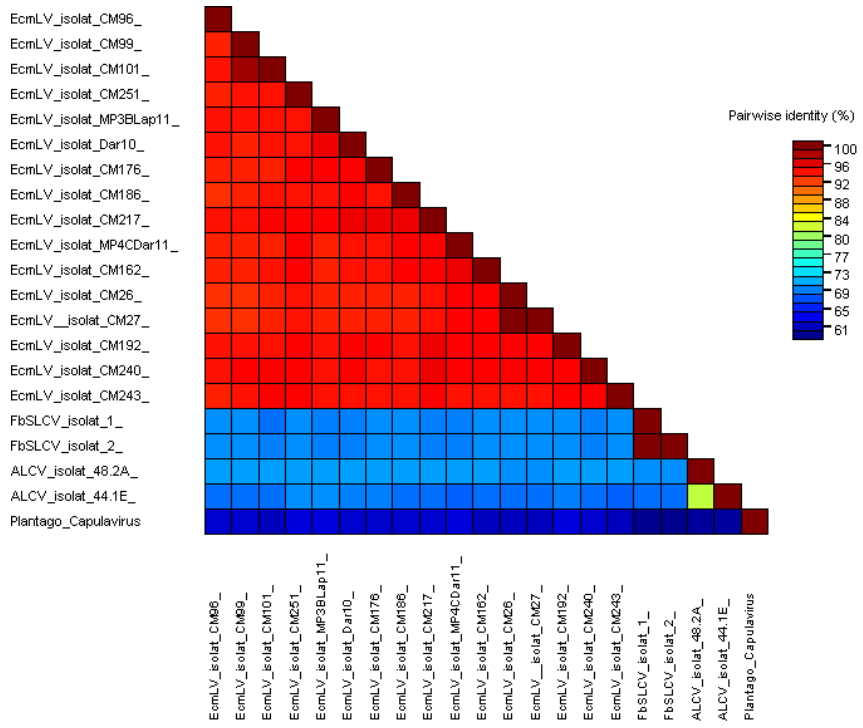


Figure 2.1: Matrice des distances génétiques moyennes par paires de génomes des capulavirus (21 génomes au total).

2.2.2.2. Démarcation des espèces de capulavirus

La comparaison de 21 génomes entiers de capulavirus deux à deux via le logiciel SDT v1.0 (Muhire *et al.*, 2013) a révélé quatre pics d'identité de séquences dans les intervalles [59 ; 64%], [67 ; 72%], [81 ; 83%] et [92 ; 97%] et trois vallées dans les intervalles [64 ; 67%], [72 ; 81%] et [83% ; 92%] (Figure 2.2). Ainsi, si nous nous basons

sur les méthodes employées pour fixer les démarcations taxonomiques à l'intérieur d'un genre de geminivirus (Varsani *et al.,* 2014b), pour les capulavirus, nous proposons de placer le seuil de l'espèce à 72% d'identité nucléotidique. Toutefois nous nous garderons de placer un seuil de démarcation pour les souches car le pourcentage d'identité des deux isolats d'ALCV [84 ; 86%] ne sont pas compris dans l'intervalle des pourcentages d'identité des isolats de EcmLV, à savoir [92 ;97%].

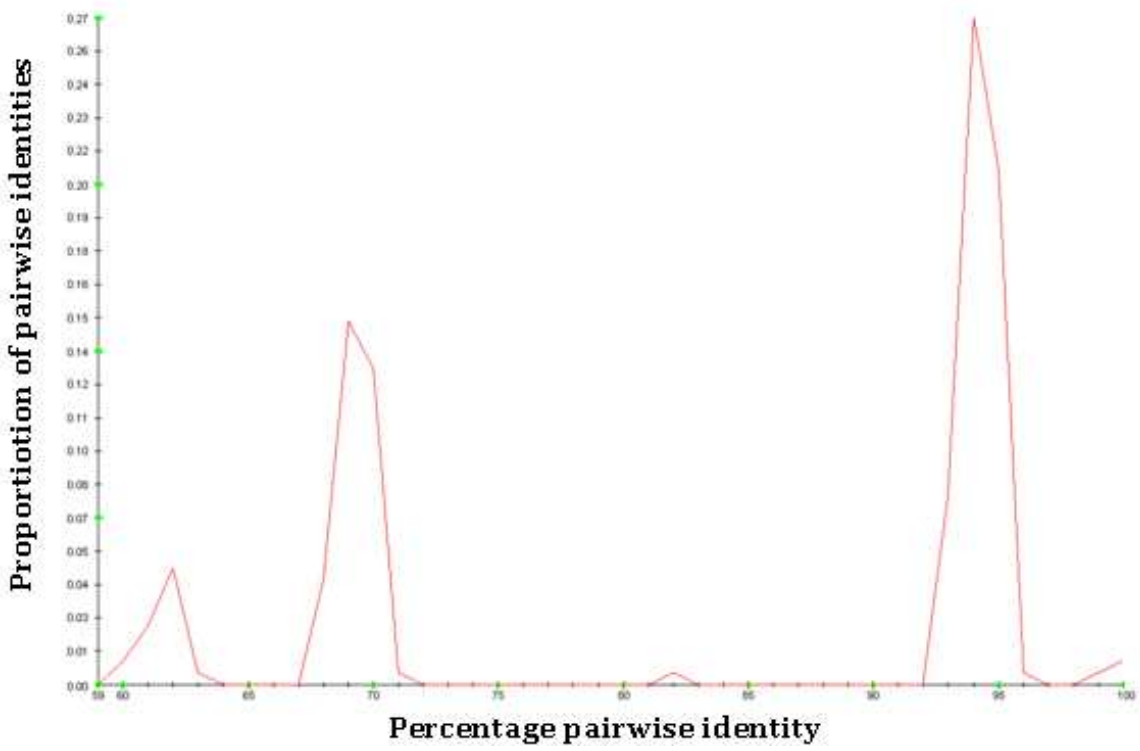


Figure 2.2 : Distribution des distances génétiques moyennes par paires de génomes des capulavirus (21 génomes) calculées par SDT v1.0 (Muhire *et al.*, 2013).

2.2.2.3. Comparaison des différents génomes de capulavirus

L'organisation des ORFs des isolats d'ALCV et de *Plantago Capulavirus* (Tableaux 2.1 et Figure 2.3) est similaire à celle d'EcmLV et FbSLCV (Bernardo *et al.* 2013). Ces génomes de capulavirus présentent des caractéristiques communes à tous les genres de la famille *Geminiviridae* tel que le gène de la *cp*. Par ailleurs, les capulavirus ont des caractéristiques qui les rapprochent des mastrevirus et des becurtovirus. C'est le cas de la présence d'un gène *rep A* (C1), d'un gène *rep* supposé épissé (C1^C2) et de la petite région intergénique SIR. Même si l'épissage reste à confirmer expérimentalement, des motifs caractéristiques d'épissage ont été détectés dans les 21 séquences du gène *rep* dont nous disposions : sites polyT, jonctions « donneur » et « accepteur » caractéristiques, et un site de « branchage ». Les capulavirus présentent aussi des ORFs qui les distinguent de tous les autres genres de *Geminivirdae*. C'est le cas de l'ORF C3, comprise dans le gène *rep*. Elle code pour une protéine putative qui n'a pas d'homologie parmi les séquences de GenBank et aucune fonction putative n'a pu lui être associée. Dans le sens du virion, plusieurs ORFs potentielles (V2, V3 et V4) pour lesquelles nous ne trouvons aucun homologue (en dehors des 3 ORFs équivalentes d'EcmLV) sur la

GenBank ont été identifiées. La présence d'un domaine transmembranaire potentiel est détecté dans toutes les protéines putatives de l'ORF V4 ce qui suggère qu'une telle protéine pourrait être impliquée dans le mouvement (MP). Par ailleurs, l'ORF V2 qui n'est portée que par EcmLV et ALCV (44.1E) présenterait également ce type de domaine. Tous les génomes étudiés, sauf celui d'ALCV (48.2A) présentent une ORF potentielle V3 pour laquelle aucune fonction n'a pu être déterminée. L'article présenté précédemment évoque l'épissage de V2 et V4 chez EcmLV. Des indices d'épissage ont également été détectés dans les ORFs chevauchantes V2 et V4 d'ALCV et *Plantago Capulavirus*. Cependant, les zones des jonctions « donneur » et « accepteur » ainsi que de « branchement » ne s'alignent pas parfaitement avec celles d'EcmLV contrairement à l'épissage de la Rep. Ici nous ne préférerons pas statuer sur un épissage quelconque entre les ORF dans le sens du virion. Une validation expérimentale sur la base de l'analyse des mARNs devrait donc être ultérieurement réalisée pour définitivement statuer sur l'organisation génomique des capulavirus dans le sens du virion.

En ce qui concerne l'origine de réplication du virion, tous ces génomes possèdent le même nonanucléotide typique des geminivirus: TAATATTAC. De plus, la structure en « tige-boucle » de ces virus contient des zones de non-hybridation (« bulles ») comme cela a déjà été décrit chez les mastrevirus.

EcmLV (2678pb)

ORF	coordonnées	nombre d'acides aminés
V1 (cp)	546-1274	242
V2 (mp?)	129-335	68
V3	248-652	134
V4 (mp?)	169-432	87
C1 (repA)	2431-1667	254
C2 (rep)	2431-1297 (intron: 1819-1681)	331
C3	2181-1816	121

FbSLCV (2771pb)

ORF	coordonnées	nombre d'acides aminés
V1 (cp)	547-1275	242
V3	249-707	152
V4 (mp?)	170-463	97
C1 (repA)	2521-1616	301
C2 (rep)	2521-1398 (intron: 1909-1749)	320
C3	2388-1906	160

ALCV 44.1E (2747pb)

ORF	coordonnées	nombre d'acides aminés
V1 (cp)	561-1295	244
V2 (mp?)	156-287	43
V3	263-715	150
V4 (mp?)	151-441	96
C1 (repA)	2537-1758	259
C2 (rep)	2537-1454 (intron: 1928-1796)	316
C3	2290-1925	121

ALCV 48.2A (2766pb)

ORF	coordonnées	nombre d'acides aminés
V1 (cp)	226-1257	343
V4 (mp?)	147-404	85
C1 (repA)	2493-1696	265
C2 (rep)	2493-1350 (intron: 1869-1734)	335
C3	2339-1866	157

Plantago Capulavirus (2832pb)

ORF	coordonnées	nombre d'acides aminés
V1 (cp)	552-1253	233
V3	260-607	115
V4 (mp?)	181-441	86
C1 (repA)	2584-1808	258
C2 (rep)	2584-1438 (intron: 1960-1822)	335
C3	2466-1957	169

Tableaux 2.1: Coordonnés des ORFs des différents capulavirus et tailles de leurs produits en acides aminés.

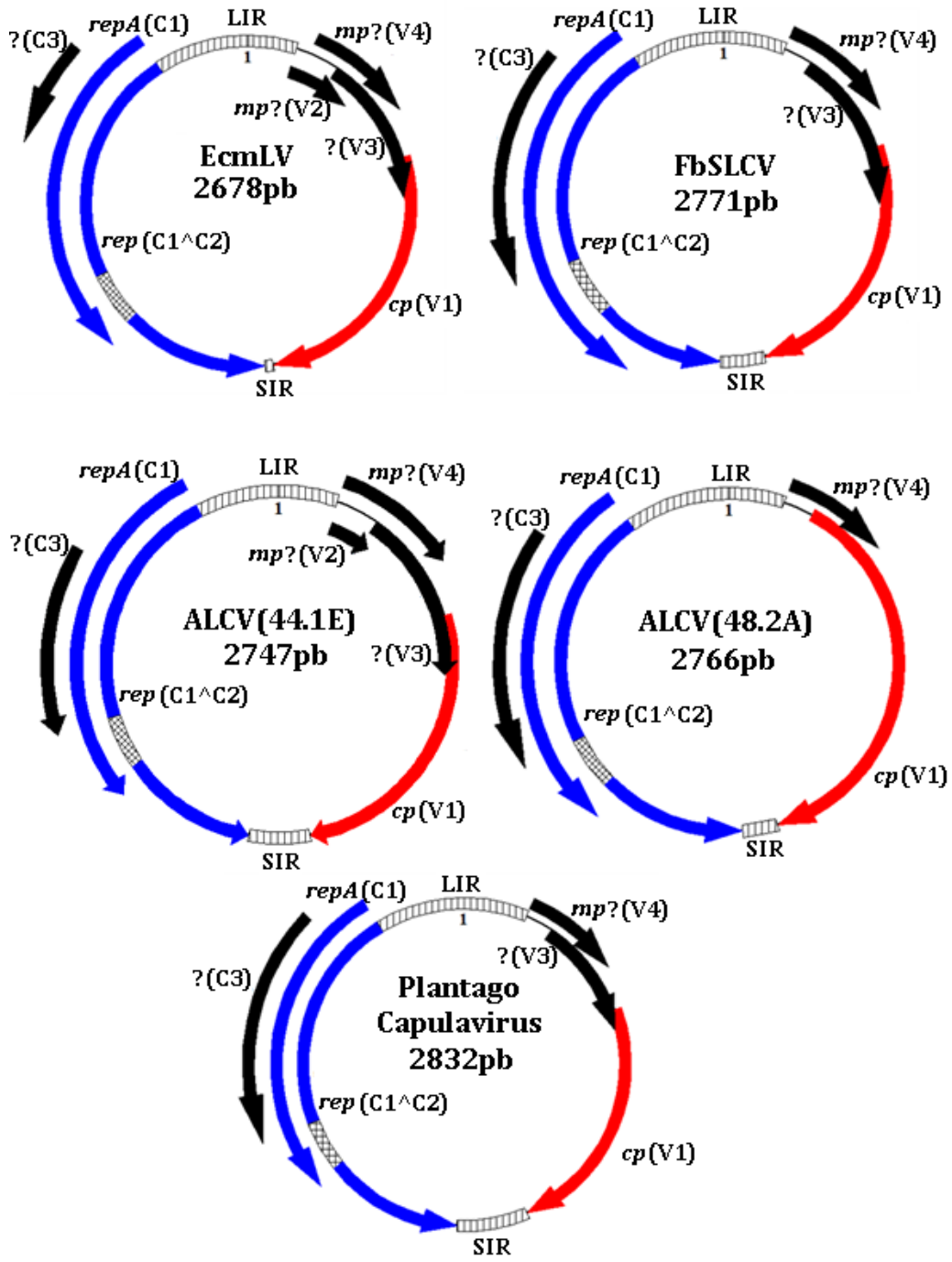


Figure 2.3 : Représentation de l'organisation génomique des différents espèces/isolats de capulavirus. Les flèches indiquent la position et l'orientation des ORFs (V= sens du virion ; C= sens complémentaire). L'intron au sein du gène *rep* est indiqué par une zone quadrillée. La SIR et la LIR sont indiqués par un hachurage. Le 1 indique l'origine de réplication du virion. *rep* et *repA* = gène de la protéine associée à la réplication, *cp* = gène de la protéine de capside, *mp* = gène de la protéine de mouvement. Les ? indiquent une absence d'homologie avec des séquences autres que celles de capulavirus sur la GenBank et donc un questionnement sur la fonction potentielle de leurs protéines putatives.

2.2.2.4. Analyse de recombinaison

L'ajout des génomes d'EcmLV et de FbSLCV aux analyses de recombinaison inter-genre chez les *Geminiviridae* nous a permis de détecter de nouveaux évènements de recombinaison au sein de cette famille (Bernardo *et al.*, 2013). L'ajout des génomes viraux provenant des échantillons de luzerne et de plantain ne nous a pas permis de détecter de nouveaux évènements de recombinaison que ce soit au sein des *Geminiviridae* ou au sein du genre Capulavirus.

2.2.2.5. Analyse phylogénétique

Nous avons inféré les relations évolutives entre les capulavirus et d'autres géminivirus représentatifs des divers genres à partir des séquences en acides aminés de la CP et de la Rep. L'arbre phylogénétique concernant la Rep contient notamment des séquences potentiellement intégrées chez le pommier, le caféier et l'igname.

Que ce soit dans l'arbre basé sur la Rep (Figure 2.4) ou l'arbre basé sur la CP (Figure 2.5) nous constatons que tous les génomes de capulavirus forment un groupe monophylétique très divergent des autres géminivirus soutenu par de fortes valeurs de bootstraps. Les arbres présentés ici ne sont pas racinés comme l'était celui réalisé sur la Rep de l'article concernant EcmLV (Bernardo *et al.*, 2013), ce qui ne nous permet pas de conclure sur les relations évolutives chez les geminivirus.

2.2.3. Conclusions et perspectives

2.2.3.1. Organisation génomique des capulavirus

Nous avons pu constater que les différents génomes de capulavirus présentaient des caractères communs aux autres geminivirus mais aussi des caractères propres. En effet, comme tous les geminivirus, ces génomes arborent une CP, une Rep et une LIR. Globalement, l'organisation des génomes des capulavirus semble se rapprocher de celle des mastrevirus, et des becurtovirus. En effet, nous avons pu détecter une SIR ainsi qu'une Rep épissée. En ce qui concerne les caractères propres aux capulavirus ils semblent tous présenter une ORF (C3) pour laquelle la fonction reste encore à identifier. De même, sur le brin viral des ORFs (V2, V3, V4) propres aux capulavirus ont été détectées et ne présentent pas d'homologues au niveau de la GenBank. Il est possible que les ORFs V2 et V4 soient associées au mouvement intraplante des capulavirus en raison d'un domaine transmembranaire potentiel. Toutefois, les ORFs n'ont été décrites que sur la base des séquences virales et leur fonctionnalité reste à être prouvée. De plus, le fait que la Rep mais aussi que les ORFs V2 et V4 puissent être sujettes à un phénomène d'épissage devrait être démontré expérimentalement. Pour cela des extractions d'ARN et des RT-PCR ciblant les transcrits devront être réalisées.

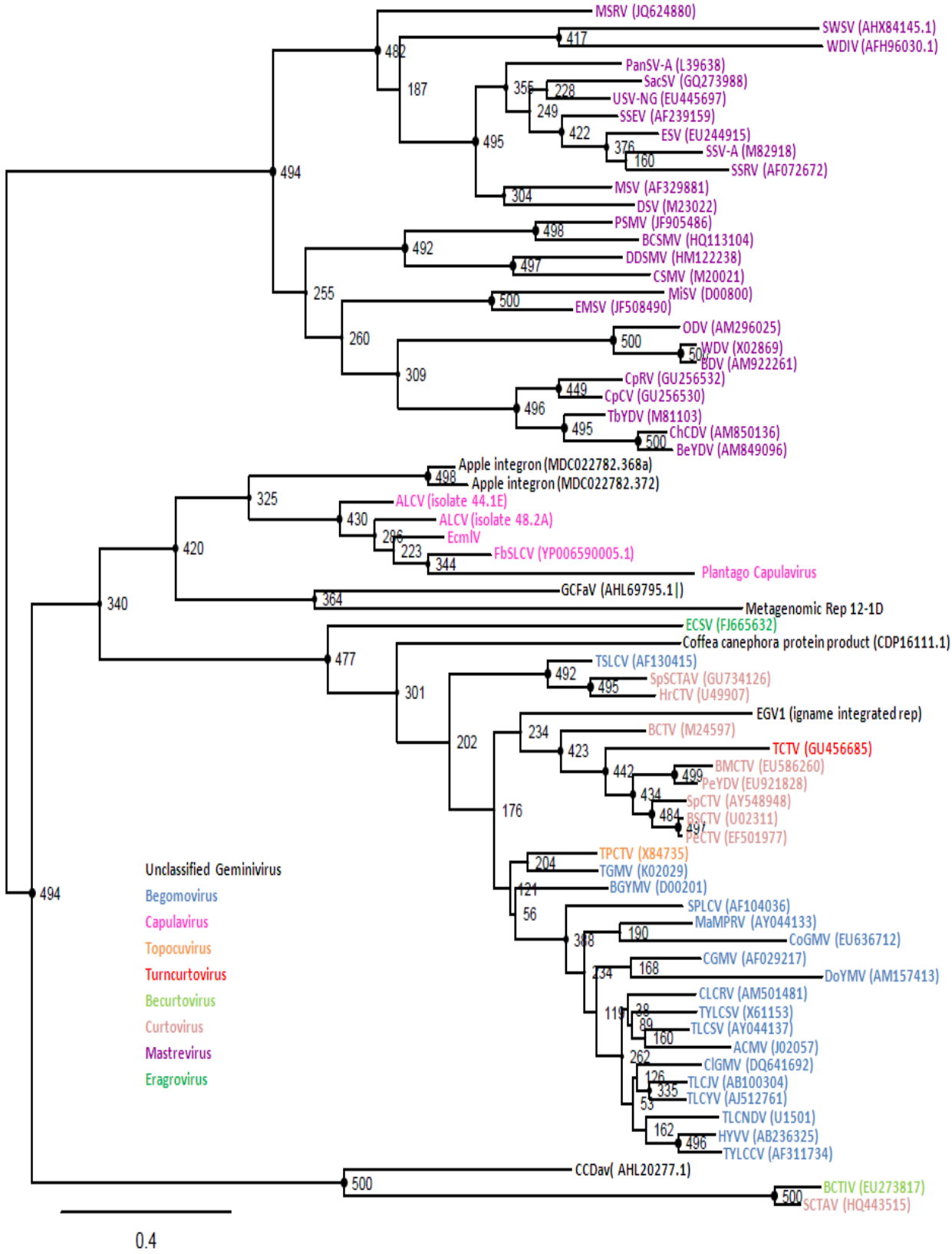


Figure 2.4 : Phylogénie réalisée selon la méthode du maximum de vraissemblance à partir de 69 séquences en acides aminées de la Rep de différents geminivirus. Cet arbre a été construit sur la base du modèle WAG+I+G. La valeur du bootstrap implémentée est de 500. A chaque séquence est attribué son genre par un code couleur. Les bootstraps sont indiqués à la base des nœuds et la taille du nœud est proportionnelle à la valeur de ce bootstrap

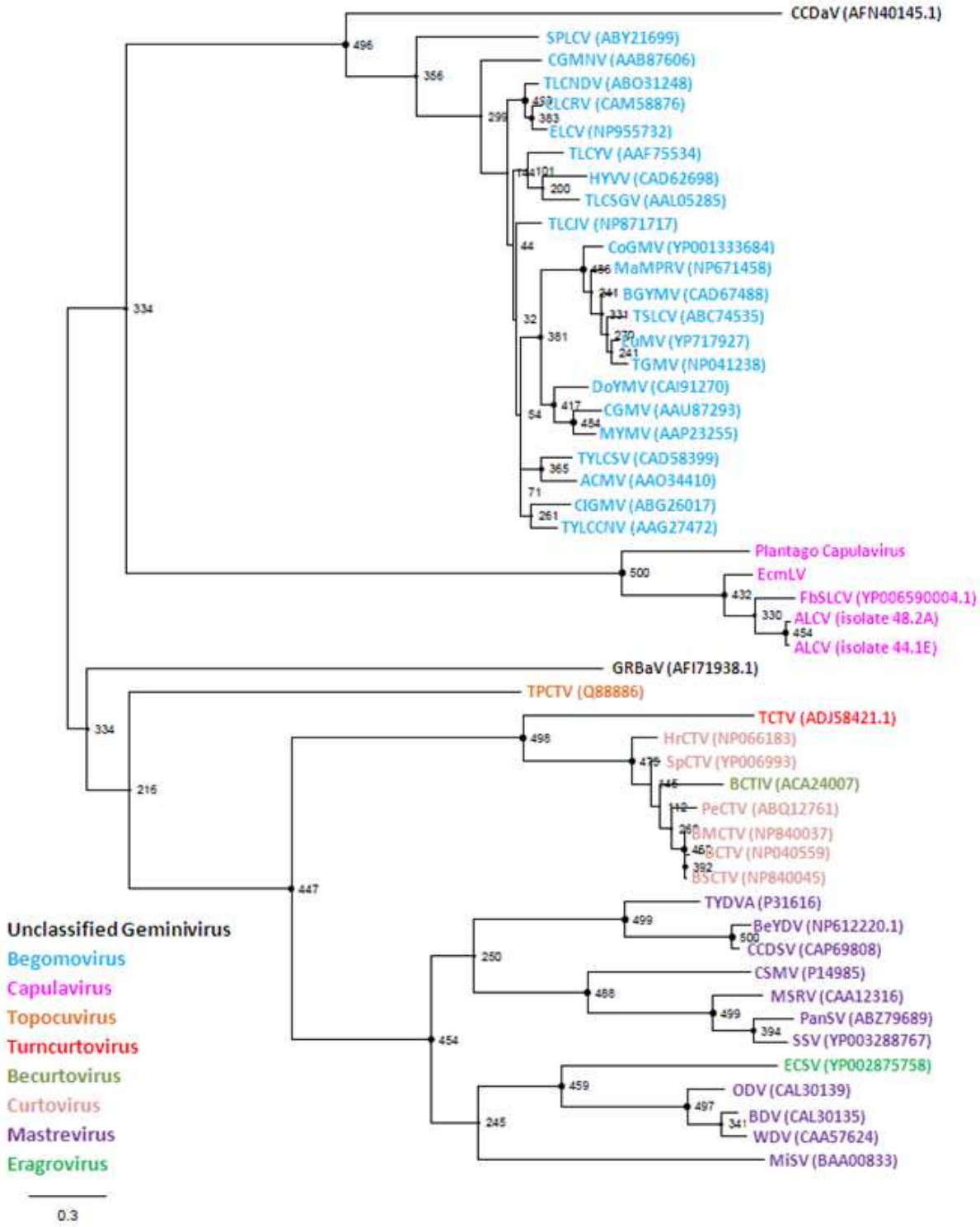


Figure 2.5 : Phylogénie réalisée selon la méthode du maximum de vraissemblance à partir de 45séquences en acides aminées de la CP de différents geminivirus. Cet arbre a été construit sur la base du modèle WAG+I+G. La valeur du bootstrap implémentée est de 500. A chaque séquence est attribué son genre par un code couleur. Les bootstraps sont indiqués à la base des nœuds et la taille du nœud est proportionnelle à la valeur de ce bootstrap

2.2.3.2. Vers un nouveau genre …

Notre étude de recombinaison nous a indiqué que les capulavirus n'étaient pas issus de phénomènes de recombinaison interspécifique et intergénériques. De plus, l'ensemble des génomes de capulavirus forme un groupe phylogénétique bien distinct des autres genres de geminivirus. Ainsi, les capulavirus peuvent être définis comme un nouveau genre de geminivirus présentant en outre des caractéristiques génomiques originales. Si la proposition qui a été faite récemment de pouvoir définir un nouveau genre sur la seule base de l'identité de séquences est finalement adoptée par l'ICTV, le nouveau groupe de geminivirus découvert dans cette thèse sera prochainement accepté comme un nouveau genre (Varsani *et al.*, 2014b). Même si la connaissance du vecteur n'est plus requise pour déposer un nouveau genre de geminivirus (exemple des Eragrovirus (Varsani *et al.*, 2014b).), la question reste brulante et importante à traiter. Il restera par ailleurs à confirmer la morphologie géminée des virions qui reste un critère apparemment incontournable pour l'adoption d'un nouveau genre. Sur la base de la comparaison deux à deux des 21 génomes disponibles, le seuil de démarcation de l'espèce au sein du genre Capulavirus a été fixé à 72% d'identité nucléotidique. Il n'est pas impossible que ce seuil puisse subir quelques ajustements sur la base d'un apport d'un plus grand nombre de séquences.

2.2.3.3. …émergent ?

Le fait que 4 espèces de capulavirus aient été décrites en l'espace de 3 ans, et qu'elles soient réparties à différents endroits distants du monde (Afrique du Sud, Finlande, France, Inde) nous amène à nous demander si nous sommes en train d'observer en temps réel une émergence virale. De plus, certains de ces virus ont été isolés à partir de plantes cultivées (luzerne, haricot) et d'autres à partir de plantes sauvages (*Euphorbia caput-medusae*, *Plantago lanceolata*). Les virus isolés sur plantes sauvages pourraient avoir des conséquences dramatiques en cas de passage sur plante cultivées. En effet, alors qu'EcmLV n'induit pas de symptômes visibles sur son hôte naturel, il cause des symptômes drastiques sur tabac et tomate. Nous avons également construit un clone potentiellement infectieux d'ALCV. Cependant, les plantes de luzerne agro-inoculées sont restées négatives à la détection du virus même un mois après inoculation. Il est possible que le clone qui a été utilisé pour la construction potentiellement agro-infectieuse soit défectif ou que la variété inoculée (var. Eugenia) ne soit pas sensible à ALCV) ; la variété sur laquelle nous avons détecté l'ALCV au champ (var. Magali) n'était pas disponible au moment des tests.

A l'heure qu'il est, nous nous demandons comment des virus aussi répandus au sein d'une famille de virus qui est scrutée par une large communauté de viroloques à travers le monde, ait pu rester dans l'ombre jusqu'à récemment. Pour ce qui est de l'EcmLV et *Plantago Capulavirus*, la réponse parait assez évidente quand on se rappelle que les virus de plantes sauvages ont été négligés pendant des années et le sont encore. Par contre, nous avons davantage de difficultés à comprendre comment des virus de haricot et de luzerne aient pu échapper à l'attention des virologues jusqu'à récemment.

Il est possible que des changements environnementaux ou agrosystémiques récents aient favorisé un saut d'hôte des plantes sauvages vers les plantes cultivées sur lesquelles se sont exprimés des symptômes sévères causés par un virus précédemment latent. Une autre raison qui est strictement technique, est le développement de nouvelles technologies telles que la métagénomique qui a considérablement augmenté le nombre d'échantillons qui peut être analysé dans une seule expérimentation et a ainsi permis de découvrir des centaines de virus comme nous l'avons montré dans la première partie de cette thèse.

Pour mieux comprendre le potentiel émergent de ces virus et le risque qu'ils représentent pour l'agriculture, nous avons cherché à évaluer leur diversité, leur prévalence, et leur mode de vection.

3. Diversité intra-spécifique et prévalence d'EcmLV et d'ALCV

3.1. Contexte et objectifs

Les pourcentages d'identité nucléotidique entre les trois génomes d'EcmLV (Dar10, Dar11 et Lap11) provenant de deux localités situées à 60km l'une de l'autre sont compris entre 93.6% et 95.3%. En supposant que l'EcmLV est un virus émergent, ces différences nucléotidiques paraissent relativement élevées au regard des faibles distances qui sépare leur lieu d'isolement. En effet, les espèces de géminivirus causant des maladies émergentes sont plutôt caractérisées par des faibles niveaux de diversité (Monjane *et al.*, 2011). La même déduction pourrait être faite pour les deux génomes d'ALCV sur luzernes qui présentent une divergence élevée (82.2%) au regard de leur éloignement géographique qui n'est que de 2km. L'analyse d'un nombre plus important d'isolats de ces deux virus devrait non seulement permettre de confirmer leur répartition géographique et leur prévalence, mais aussi de nous aiguiller concernant leur mode de transmission d'après la structuration géographique de leur diversité.

3.2. Matériel et méthodes

3.2.1. Echantillonnage à l'aveugle

Les échantillonnages détaillés ci-dessous ont été effectués à l'aveugle, c'est-à-dire sans tenir compte des symptômes que les plantes pouvaient arborer. Chacun des échantillons a été géo-référencé, conservé à 4°C pendant les campagnes d'échantillonnage puis à -80°C au laboratoire.

De septembre à octobre 2012, nous avons réalisé une campagne d'échantillonnage dans huit localités distinctes de la région du Western Cape en Afrique du Sud (Figure 2.6). Nous avons récolté un total de 301 nouveaux échantillons d'*E. caput-medusae* dans 8 localités (Tableau 2.2a). Si l'on tient compte des échantillons que nous avions prélevés en 2010 et en 2011 (voir l'article présenté dans ce chapitre), nous avons un total de 311 échantillons à notre disposition. En mars 2014 nous avons récolté

202 échantillons de luzerne au sein de 8 localités différentes proche de la commune d'Arles (Camargue) (Figure 2.7). Les parcelles sur lesquelles ces luzernes ont été prélevées ont été renommées en fonction de la commune, du nom de l'exploitant, ou du centre de recherche (Tour du Valat) (Tableau 2.2b). Durant le mois d'août 2014, nous avons également collecté des luzernes dans des parcelles près des communes de Montferrier (Hérault), La Salvetat-sur-Agoût (Hérault) et Brassac (Tarn) (Figure 2.7). Nous disposons donc d'un total de 223 nouveaux échantillons de luzerne (Tableau 2.2b).

Figure 2.6 : Carte représentant les différentes zones d'échantillonnage d'*Euphorbia caput-medusae* dans la région du Western Cape. Les 8 zones d'échantillonnages sont indiquées par une punaise rouge.

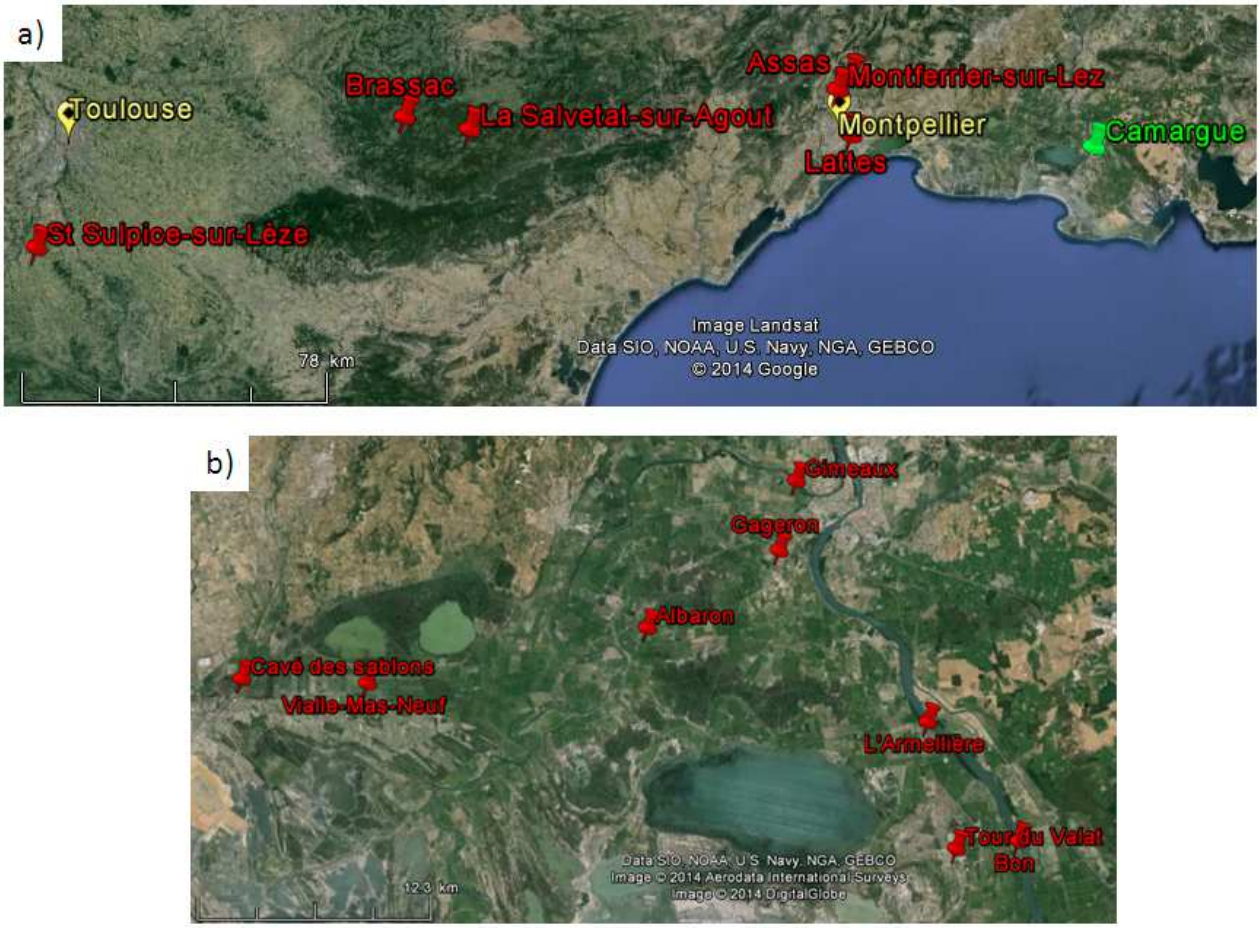


Figure 2.7 : Cartes représentant les différentes zones d'échantillonnage de luzerne dans le Sud de la France. a) Carte représentant les différentes zones d'échantillonnage à l'échelle du Sud de la France, les zones d'échantillonnages en Camargue (punaise verte) sont détaillées dans la carte **b) Carte représentant les différentes zones d'échantillonnage en Camargue**, les 14 zones d'échantillonnages sont indiquées par une punaise rouge.

a)

Localité	Nombre d'échantillons	Echantillons positifs	Prévalence (%)
Buffelsfontein	172	35	20
Atlantis	21	0	0
Silver Stream	17	1	6
Laaiplek	16	1	6
Pater Noster	20	7	35
Yzerfontein centre	20	0	0
Yzerfontein beach	20	0	0
Lion's head	15	0	0
TOTAL	**301**	**44**	**15**

b)

	Localité	Nombre d'échantillons	Echantillons positifs	Prévalence (%)
CAMARGUE	Gimeaux	24	11	46
	Albaron	17	6	35
	Bon	24	6	25
	Vialle-Mas-Neuf	15	2	13
	Gageron	15	2	13
	L'Armellière	34	1	3
	Cavé des sablons	3	0	0
	Tour du Valat	70	0	0
AUTRES	Brassac	5	0	0
	La Salvetat-sur-Agout	9	2	22
	Montferrier-sur-Lez	7	1	14
	TOTAL	**223**	**31**	**14**

Tableau 2.2 : Tableaux récapitulant les prévalences virales globales et locales. a) Prévalences de EcmLV dans la région du Western Cape en 2012. b) Prévalences de ALCV dans le Sud de la France en 2014.

3.2.2. Echantillonnage de luzernes symptomatiques

Suite à la détection d'ALCV dans des plants de luzerne échantillonnés à l'aveugle, nous avons émis des suppositions quant aux symptômes induits par ALCV. En effet, les plants positifs à la détection d'ALCV étaient tous chétifs ; leur feuilles étaient incurvées, gaufrées et jaunies. Afin de pouvoir confirmer l'association de ces symptômes à l'infection par ALCV, nous avons effectué un échantillonnage de luzerne au terrain sur la base de ces symptômes. Ainsi, nous avons collecté 11 échantillons près de la commune de Saint-Sulpice-sur-lèze (Haute-Garonne), 9 à Lattes (Hérault), 3 à Assas (Hérault) et 20 à Gimeaux (Camargue) (Figure 2.7). De plus, nous avons collecté 5 plantes non-symptomatiques à Gimeaux et 5 autres à Albaron (Figure 2.7). Nous disposions donc au total de 43 échantillons symptomatiques et de 10 non-symptomatiques.

3.2.3. Extractions d'ADN

La technique utilisée concernant les extractions d'ADN à partir des échantillons végétaux est celle décrite part Bernardo *et al.* (Bernardo *et al.*, 2013).

3.2.4. Détection virale par PCR et séquençage de zones d'intérêt

Les extraits d'ADN issus des différents échantillons d'*Euphorbia caput-medusae* ont été soumis à une PCR avec les amorces EcmLV-136F et EcmLV-730R (Bernardo *et al.*, 2013) afin de détecter la présence d'EcmLV. Les échantillons positifs suite à cette PCR ont été soumis à une amplification via une deuxième paire d'amorces : EcmLV-1775Fet EcmLV-2433R (Annexe 7) afin de pouvoir réaliser une étude phylogénétique sur une partie du gène *rep*.

Les extraits d'ADN provenant des différents échantillons de luzerne ont été soumis à une amplification PCR avec les amorces Luz-CP-F et Luz-CP-R (Annexe 7). Ce couple d'amorces permet à la fois de détecter ALCV et d'amplifier une partie du gène codant pour la *cp* nous permettant ainsi de réaliser une phylogénie sur une partie de ce gène.

La visualisation des résultats PCR a été réalisée via une électrophorèse sur gel d'agarose à 1% (100V, 20min). Les produits amplifiés concernant la *rep* des isolats d'EcmLV et la *cp* des isolats d'ALCV ont été séquencés via la technique Sanger par la société Beckman Coulter Genomics.

3.2.5. Obtention de génomes entiers de divers isolats d'EcmLV par amplification via des amorces chevauchantes, clonage et séquençage.

Afin d'obtenir les génomes entiers des différents isolats d'EcmLV nous avons réalisé une RCA sur les extraits d'ADN des échantillons d'*E. caput-medusae* détectés positifs pour la présence d'EcmLV. Puis nous avons réalisé une PCR sur le produit RCA à l'aide des amorces chevauchantes Dar-1981F et Dar-1966R. Les fragments PCR ont ensuite été clonés et séquencés. L'ensemble du protocole est décrit dans Bernardo *et al.* 2013. Concernant les isolats d'ALCV, l'obtention des génomes entiers est en cours.

3.2.6. Calcul de la prévalence d'EcmLV et d'ALCV

La formule suivante a été utilisée :

$$Prévalence(\%) = \frac{nombre\ de\ plants\ infectés\ par\ localité}{nombre\ de\ plants\ testés\ par\ localité} \text{x}100$$

3.2.7. Analyse du polymorphisme des séquences nucléotidiques des différents isolats d'EcmLV et d'ALCV

Les séquences issues des PCR effectuées sur une partie de la *rep* pour EcmLV et sur une partie de la *cp* pour ALCV ont été coupées (« trimmées ») via le logiciel Geneious 5.4.6 (Kearse *et al.*, 2012) puis alignées via la méthode ClustalW (Thompson *et al.*, 1994) implémentée dans MEGA 5.2.1 (Tamura *et al.*, 2011). Le logiciel MEGA5.2.1 nous a également permis de générer une matrice des distances de similarité des génomes deux à deux (pairwise distance matrix).

3.2.8. Analyse phylogénétique des différents isolats d'EcmLV et d'ALCV et détection d'évènements de recombinaison

Des analyses phylogénétiques par maximum de vraisemblance ont été réalisées grâce au logiciel PhyML 3.1 (Guindon *et al.*, 2005) en ayant sélectionné au préalable le meilleur modèle évolutif décrivant nos alignements via le logiciel JModelTest 2.1.6 (Posada, 2008). Les analyses ont été faites sur des alignements effectués sur une partie de la *rep* d'EcmLV (509pb), sur une partie de la *cp* d'ALCV (423pb) et sur 16 séquences complètes de la *cp* et de la *rep* d'EcmLV. L'arbre phylogénétique a été construit après réalisation de 500 bootstraps.

Nous avons cherché si des évènements de recombinaison avaient eu lieu entre les 16 génomes complets d'EcmLV via le logiciel RDP3.44 (Martin *et al.*, 2010). La méthode est décrite dans Bernardo *et al.* (Bernardo *et al.*, 2013).

3.2.9. Corrélation entre distances génétiques et géographiques des isolats

Un test de Mantel (Mantel, 1967) a été réalisé afin d'évaluer la corrélation entre distances génétiques et distances géographiques d'une part pour les différents isolats d'EcmLV à l'échelle régionale (Western Cape) et locale (Bufflesfontein) et d'autre par pour les différents isolats d'ALCV à l'échelle de la Camargue puis du Sud de la France. Ce test a été réalisé à l'aide du logiciel XLSTAT à partir des matrices de distances génétiques obtenues précédemment grâce au logiciel MEGA 5.2.1 et des matrices de distances géographiques obtenues grâce au logiciel Geographic Distance Matrix Simulator 1.23 à partir des coordonnées GPS des échantillons. Le test est basé sur 10 000 permutations et sur un test de corrélation de Pearson, avec un seuil α=5%.

3.3. Résultats

3.3.1. Prévalence d'EcmLV dans la région du Western Cape et d'ALCV dans le Sud de la France

Sur les 301 échantillons d'*E. caput-medusae* récoltés en 2012 nous avons obtenu un total de 44 plantes positives à la détection d'EcmLV. EcmLV est présent sur 4 des 8 localités dans lesquelles nous avons réalisé l'échantillonnage et la prévalence associée à chacune de ces 4 localités varie de 6% (Silver Stream, Laaiplek) à 35% (Pater Noster). La prévalence d'EcmLV sur la totalité des localités échantillonnées atteint 15% (Tableau 2.2a).

Sur les 223 échantillons de luzerne récoltés à l'aveugle en 2014, nous avons obtenu un total de 31 plantes positives à la détection d'ALCV. ALCV est présent sur 6 des 8 localités échantillonnées en Camargue avec des prévalences allant de 3% (L'Armellière) à 46% (Gimeaux) et une prévalence moyenne sur les 8 localités de 14%. Si l'on tient compte de l'échantillonnage complet (Camargue, Brassac, La Salvetat-sur-Agout et Montferrier) la prévalence d'ALCV est également de 14%. (Tableau 2.2b).

3.3.2. Confirmation des symptômes causés par ALCV

Nous avons détecté ALCV sur 39 des 43 échantillons de luzerne collectés sur la base de symptômes (91%). Nous n'avons pas détecté la présence d'ALCV sur les 10 plants non-symptomatiques. Sur cette base nous pouvons fortement présumer qu'ALCV induit une incurvation, un gaufrage et un jaunissement des feuilles chez *M. sativa* ainsi qu'un retard ou un arrêt de sa croissance (Annexe 8).

3.3.3. Analyse du polymorphisme des séquences nucléotidique des isolats d'EcmLV et d'ALCV

Une moyenne de 4,3% de distance génétique par paire moyenne (average pairwise genetic distance) avec une amplitude de 0% à 7,1%, ont été obtenues à partir des 44 séquences de 509pb. Le clonage nous a permis d'obtenir 16 génomes entiers (12 issus de Bufflesfotein, 3 de Pater Noster et 1 de Laiiplek). Ces génomes entiers ont une moyenne de 5,8% de distance génétique par paire moyenne (avec un minimum de 1,5% et un maximum de 7,3%). Si l'on considère maintenant les variations génétiques au niveau de leur *repA* ils ont une moyenne de 4,5% de distance génétique par paire moyenne (minimum de 0% et maximum de 6%) alors que pour leur *cp,* la moyenne des distances génétiques par paire moyenne est égale à 6,1% (minimum de 0,3% et maximum de 9,4%).

Nous disposons au total de 51 isolats d'ALCV à partir desquels nous avons pu obtenir une partie du gène codant pour la CP (423pb). Ces séquences alignées nous ont permis de déterminer qu'elles avaient une moyenne des distances génétiques par paire moyenne de 3,5% avec un minimum de 0% et un maximum de 9%.

3.3.4. Analyse phylogénétique des isolats d'EcmLV et d'ALCV

L'arbre phylogénétique réalisé selon la méthode de maximum de vraisemblance à partir des 44 séquences partielles du gène codant pour la Rep (509pb) est représenté sur la Figure 2.8. On peut remarquer que bien que les isolats d'EcmlV de Pater Noster se regroupent dans un même groupe phylogénétique en haut de l'arbre (en association avec certains isolats de Bufflesfontein et de Silver Stream) et que la majorité des isolats d'EcmlV de Bufflesfontein sont regroupés dans un même groupe phylogénétique en bas de l'arbre, il ne semble pas y avoir de structuration géographique des populations d'EcmLV en fonction des localités étant donné que les isolats de Bufflesfontein sont répartis dans la totalité de l'arbre.

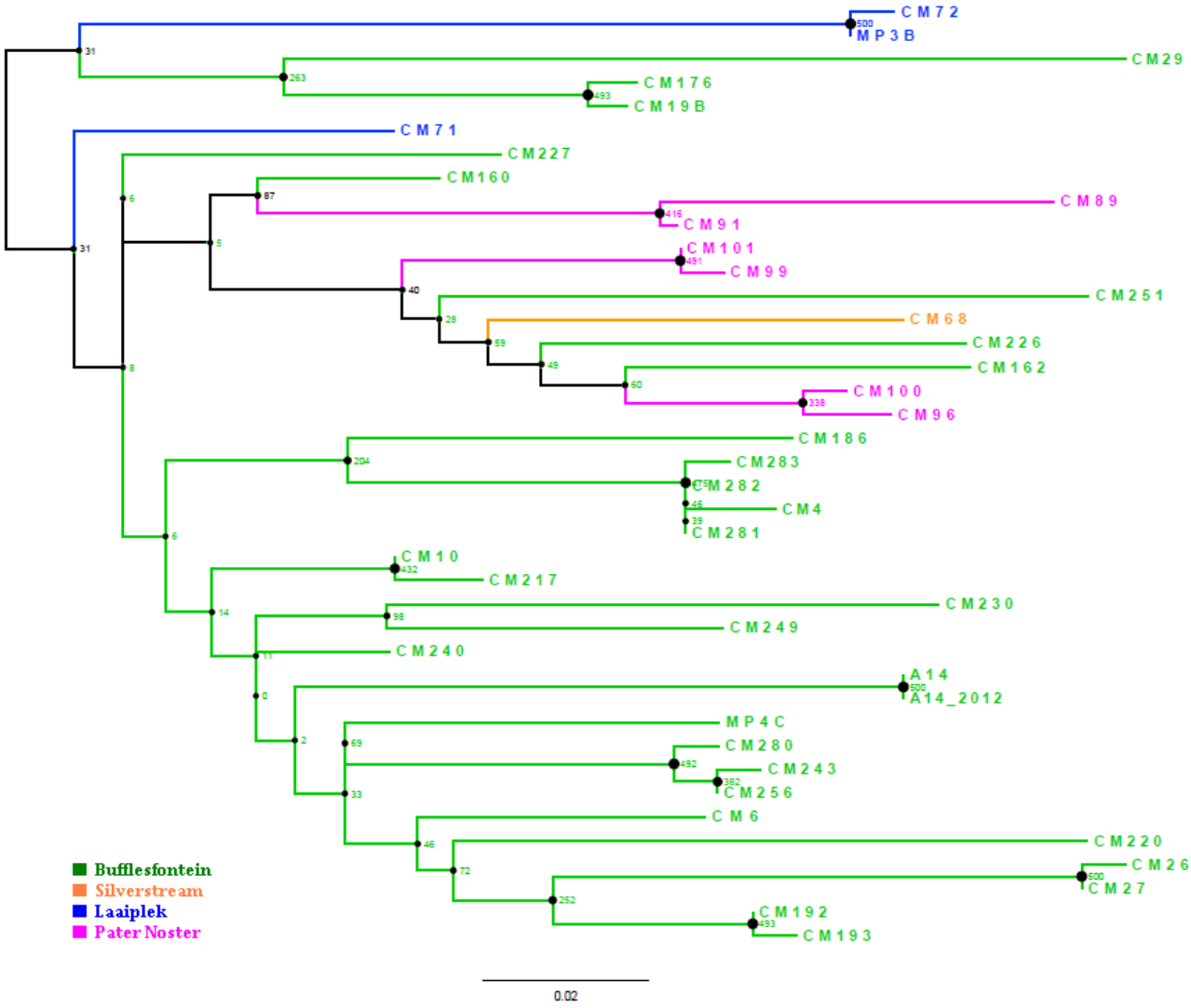


Figure 2.8 : Phylogénie réalisée selon la méthode du maximum de vraisemblance à partir de 40 séquences partielles de la *rep* des isolats d'EcmLV (509pb). Cet arbre a été construit sur la base du modèle GTR+I+G. Le nombre de bootstraps implémentés est de 500. A chaque échantillon est attribuée sa localité par un code couleur. Les valeurs des bootstraps sont indiquées à la base des nœuds et la taille du nœud est proportionnelle aux valeurs des bootstraps.

L'arbre phylogénétique réalisé à partir d'une partie de la séquence du gène codant pour la CP (423pb) des isolats d'ALCV (Figure 2.9) nous indique qu'il ne semble pas y avoir non plus de structuration génétique en fonction des localités françaises. En effet, on voit que les isolats des différentes localités ne forment pas de groupes monophylétiques clairs et distincts et qu'ils sont dispersés tout au long de l'arbre. On constate également la présence d'un groupe phylogénétique présentant une faible diversité génétique regroupant des isolats provenant de Camargue, de St Sulpice-sur-lèze et de La Salvetat, alors que le reste des isolats hors de ce groupe semblent plus diversifiés.

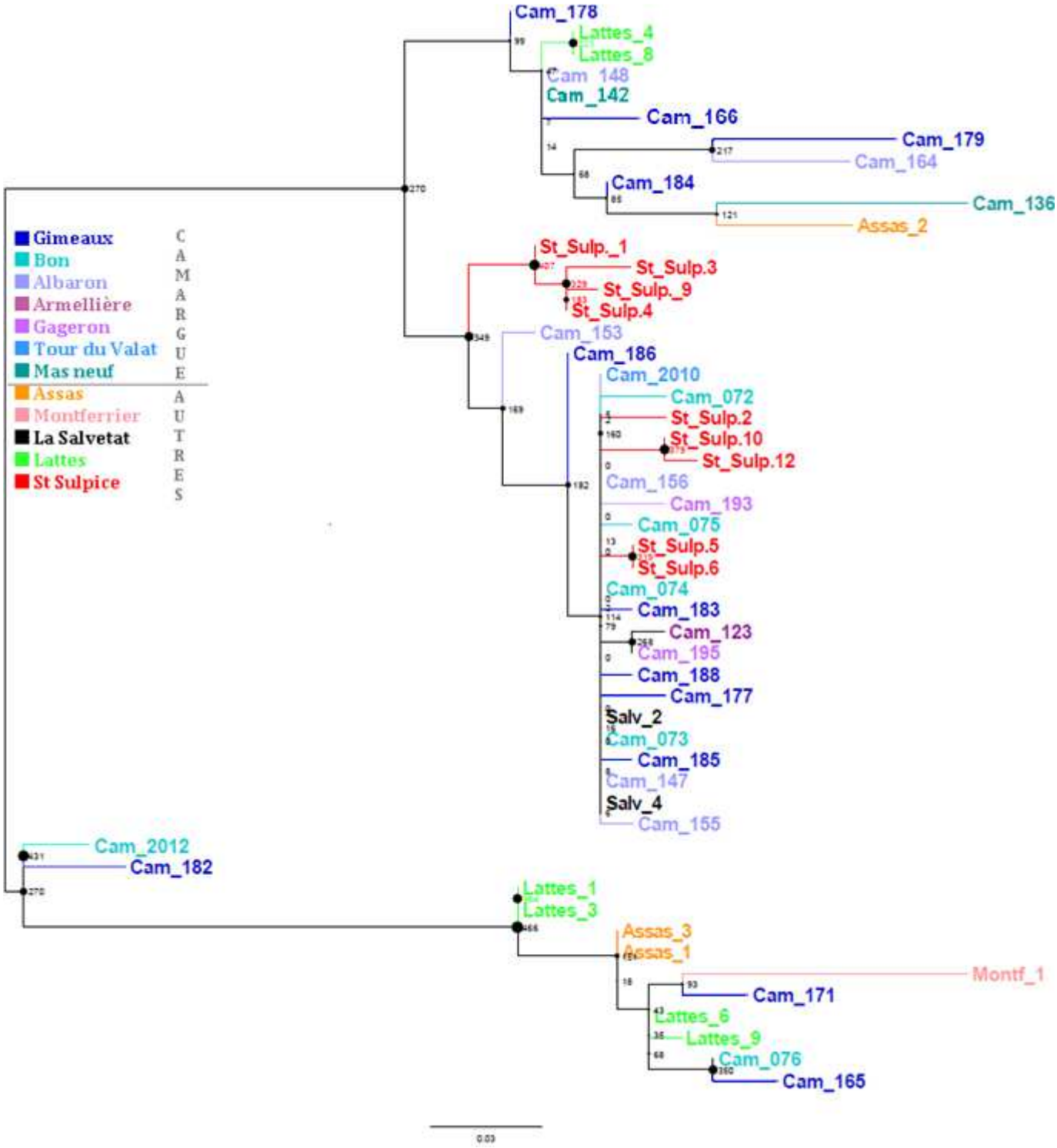


Figure 2.9 : Phylogénie réalisée selon la méthode du maximum de vraisemblance à partir de 51 séquences partielles de la *cp* des isolats d'ALCV (423pb). Cet arbre a été construit sur la base du modèle HKY+I+G avec 500 bootstraps. A chaque échantillon est attribué sa localité par un code couleur. Les valeurs des bootstraps sont indiquées à la base des nœuds et la taille du nœud est proportionnelle aux valeurs des bootstraps.

3.3.5. Analyse de la recombinaison au sein des isolats d'EcmLV

Des arbres phylogénétiques réalisés à partir de séquences totales de la *rep* et de la *cp* d'EcmLV semblent nous indiquer que des évènements de recombinaison ont pu avoir lieu (Figure 2.10). En effet ces arbres ne sont pas congruents. Notre analyse de recombinaison indique que plusieurs isolats semblent effectivement être issus d'évènements de recombinaison (Tableau 2.3). Des isolats génétiquement proches de CM280 et A14 provenant de Bufflesfontein ont « donné naissance » à au moins quatre isolats (MP4C, CM26, CM27 et CM162). De plus, il semblerait que CM96 soit issu de parents provenant de localités différentes. Les évènements de recombinaison ont été détectés au niveau de diverses régions génomiques avec des fragments plus ou moins longs (de 187 à 1544pb). Les évènements détectés sont cohérents avec les phylogénies non congruentes de la *cp* et de la *rep*.

Isolat recombinant Localité	Parent majeur Localité	Parent mineur Localité	Taille du fragment recombiné (pb)	position du fragment recombiné	Méthodes de détection
MP4C Bufflesfontein	CM280 Bufflesfontein	A14 Bufflesfontein	187	*mp* potentielle	4
CM26-CM27 (clonales) Bufflesfontein	CM280 Bufflesfontein	A14 Bufflesfontein	298	*mp* potentielle	4
CM162 Bufflesfontein	CM280 Bufflesfontein	A14 Bufflesfontein	223	*mp* potentielle	4
CM96 Pater Noster	CM192 Bufflesfontein	CM27-CM26 Bufflesfontein	1544	*rep* et *cp*	6
	MP3B Laaiplek	CM162 Bufflesfontein	666	*rep*	4

Tableau 2. 3 : Tableau récapitulatif des évènements de recombinaison ayant eu lieu entre divers isolats d'EcmLV. L'analyse a été inférée à partir de génomes entiers. La colonne « Méthodes de détection » indique le nombre de méthodes ayant détecté une recombinaison avec une p-value<0.05, nous n'avons gardé que les isolats pour lesquels au moins 4 méthodes ont détecté un évènement de recombinaison avec une p-value significative.

3.3.6. Analyse des corrélations entre distances génétiques et distances géographiques des isolats d'EcmLV et d'ALCV

D'après les tests de Mantel, il semblerait ne pas y avoir de corrélation entre les distances génétiques et géographiques des isolats d'ALCV que ce soit au niveau local (Camargue, p-value=0.615) ou au niveau régional (Sud de la France, p-value=0.230). Pour EcmLV, ces tests ont montré que les distances génétiques et géographiques des différents isolats n'étaient pas corrélées au niveau local (Bufflesfontein, p-value=0.086) mais l'étaient au niveau régional (Western-Cape, p-value=0.002).

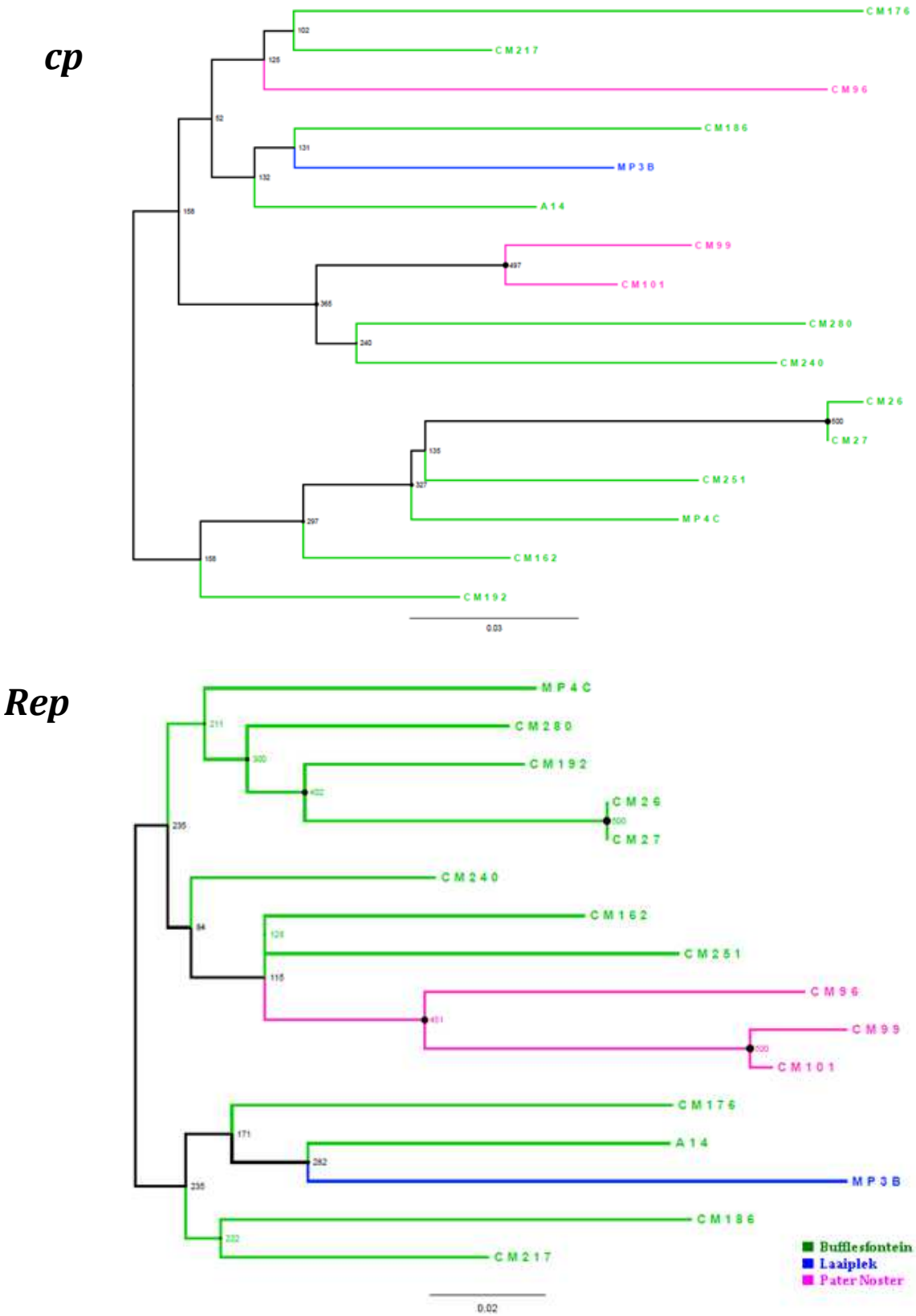


Figure 2.10 : Phylogénies réalisées selon la méthode du maximum de vraissemblance à partir de séquences entières de la *cp* et de la *rep* de 16 isolats d'EcmLV. Ces arbres ont été construits sur la base du modèle GTR+I+G avec 500 bootstraps. A chaque échantillon est attribuée sa localité par un code couleur. Les bootstraps sont indiqués à la base des nœuds et la taille du nœud est proportionnelle aux valeurs des bootstraps.

3.4. Discussion

3.4.1. Répartitions et prévalences d'EcmLV et d'ALCV

Notre étude montre qu'EcmLV a été détecté de Laiiplek à Silverstream sur une distance de 60km dans la région du Western Cape. ALCV a quant à lui été détecté sur 270 kms dans le Sud de la France, du Midi Toulousain jusqu'en Camargue. Etant donné que les virus ont été trouvés dans la majorité des sites échantillonnée, on peut supposer que l'aire géographique de la présence de ces virus est beaucoup plus vaste.

Bien que la prévalence d'EcmLV sur la totalité des localités échantillonnées atteigne 15%, la prévalence à l'intérieur de chaque localité varie de 0% à 30%. D'autre part, la prévalence d'ALCV atteint 14% en Camargue avec des prévalences intra-localité très variables. Les prévalences d'ALCV au niveau de la Camargue et d'EcmLV au niveau du Western Cape sont proches de celle reportée pour le *Bean golden mosaic virus* (BGMV) qui touche les cultures de haricots verts en Argentine (Alemandri, 2012). Les variations de prévalences inter-localités peuvent être dues à des différences d'efforts d'échantillonnage. En effet, en ce qui concerne la recherche d'EcmLV, un effort considérable a été fourni sur la localité de Bufflesfontein (177 échantillons récoltés) alors que nous n'avons récolté que 15 à 23 échantillons sur les autres localités. De même pour ALCV, un effort considérable a été fourni en Camargue alors qu'au lieu-dit « Cavé des sablons » et dans les localités extérieures à la Camargue les échantillonnages ont été relativement modestes.

3.4.2. Symptômes causés par ALCV sur la luzerne cultivée

Nous avons pu déterminer que 91% des plants collectés sur la base de symptômes (plants chétifs, feuilles jaunies, gaufrées et incurvées) étaient infectés par ALCV. Ainsi nous supposons fortement que ces symptômes sont causés par ALCV. Cependant il est nécessaire de vérifier les postulats de Koch afin de bien démontrer l'association entre les symptômes et l'infection virale. Par conséquent, il sera nécessaire de réaliser un clone infectieux à partir d'un isolat d'ALCV et d'effectuer des agroinoculations sur des luzernes saines afin de confirmer la symptomatologie liée à ALCV tout comme nous l'avons fait pour EcmLV (Bernardo *et al.*, 2013).

3.4.3. Diversité génétique d'EcmLV et d'ALCV

Les niveaux de diversité génétique d'ALCV et d'EcmLV (3,5-4.3%) sont légèrement supérieurs à ceux reportés pour certains géminivirus émergents. Par exemple, les différents isolats du MSV ont un maximum de 2.6% de divergence à l'échelle du continent Africain (Monjane *et al.*, 2011). Cependant ce niveau de diversité génétique élevé est du même ordre que ceux reportés pour le *Tomato leaf curl virus* qui infecte *Solanum nigrum* dans les zones sauvages d'Espagne ou de ceux des virus infectant les plantes sauvages du genre *Eupatorium* au Japon (Garcia-Andres *et al.*, 2006; OOi *et al.*, 1997). Il faut cependant noter que *E. caput-medusae* et *Medicago sativa* sont des plantes pérennes, ce qui peut jouer un rôle dans la conservation de la diversité par

une pression de sélection plus faible que chez les plantes annuelles qui obligent le virus a être compétitif pour le passage de plante à plante.

Par ailleurs, des études phénotypiques concernant *E. caput-medusae* indiquent qu'elle représenterait un complexe d'espèces (Williamson, 2011), une observation qui devrait être testée par des études génétiques. Si cette diversité phénotypique était reliée à une diversité génotypique, il faudrait voir dans quelle mesure la diversité génétique des plants d'*E. caput-medusae* pourrait être corrélée à la diversité génétique d'EcmLV.. Il serait donc intéressant de réaliser des arbres phylogénétiques à partir de marqueurs propres à ces euphorbes tels que les séquences ITS (Steinmann and Porter, 2002) ou chloroplastiques (Taberlet *et al.*, 1991) afin d'analyser la congruence entre les phylogénies des plantes hôtes et des isolats viraux.

3.4.4. Relations entre distances génétiques et distances géographiques entre les différents isolats d'EcmLV et d'ALCV

Les phylogénies inférées sur une portion de génome des isolats d'EcmLV et d'ALCV semblent nous indiquer que ces espèces de géminivirus ne sont pas soumises à des phénomènes d'isolement génétique à l'échelle locale. De plus, à l'échelle régionale, ALCV ne subit pas non plus d'isolement génétique par la distance. Ainsi, bien que les isolats de St Sulpice soient distants de centaines de kilomètres des isolats de Camargue ils sont génétiquement très proches de ces derniers. A contrario, le test de Mantel effectué sur les isolats d'EcmLV au niveau régional indique une structuration géographique des isolats. Ce résultat doit être considéré avec prudence car il pourrait s'expliquer par un biais d'échantillonnage ; comme évoqué ci-dessus, la taille de l'échantillon de Bufflesfontein est disproportionnée par rapport à celui des autres localités. Le fait qu'on ne retrouve pas de correspondance entre les distances génétiques et les distances géographiques de manière générale nous amène à supposer que les isolats de ces virus sont continuellement brassés à l'échelle régionale. Ce phénomène pourrait être expliqué par un mode de transmission efficace des virus à l'aide d'un vecteur qu'il reste à découvrir. En effet, des virus transmis par des insectes vecteurs tels que les geminivirus peuvent être dispersés sur de grandes distances, ce qui a pour conséquence de brouiller à une certaine distance la structuration géographique des populations virales par flux de gènes. Cependant, d'autres phénomènes biologiques pourraient expliquer ce manque de structuration géographique. En effet, si on s'intéresse au mode de reproduction d'*E. caput-medusae*, on constate qu'elle est capable de se reproduire de façon végétative ce qui permettrait de transmettre le virus à la descendance. Cette hypothèse pourrait être testée au laboratoire car nous possédons des plants infectés d'EcmLV. De plus, nous avons constaté que les plants récoltés étaient assez souvent broutés par un herbivore présent dans la région du Cap (Steenbok-*Raphicerus campestris*), ce qui représenterait un mode de transmission inédit dans la famille des *Geminiviridae*. Une généralisation de cette hypothèse peut être raisonnablement écartée pour ALCV car la configuration spatiale du paysage français rend difficilement envisageable le déplacement d'un herbivore du même type sur plus

de 200km ! Pour ALCV, une hypothèse classique d'un insecte vecteur piqueur-suceur nous semble davantage pertinente.

3.4.5. Recombinaison intraspécifique chez EcmLV

La facilité avec laquelle des recombinant ont été détectés sur un petit échantillons de génomes complets d'EcmLV est cohérente avec la forte propension des geminivirus à la recombinaison intraspécifique (Padidam *et al.*, 1999; Rybicki, 1994). Nous avons pu détecter des évènements de recombinaison au sein d'isolats provenant d'une même localité mais également au sein d'isolats provenant de localités différentes ce qui renforce l'idée d'une transmission à longue distance des virus. Ces phénomènes de recombinaison pourraient également contribuer à masquer la structuration géographique de populations virales par un brassage efficace de leur génome. Il serait intéressant d'obtenir d'autres génomes entiers d'EcmLV pour mesurer l'étendue de ce phénomène au sein de l'espèce. De plus, l'obtention des génomes entiers d'ALCV pourra également nous permettre d'effectuer le même type d'analyse.

4. Etude de la transmission des capulavirus

4.1. Contexte et objectifs

Comme évoqué précédemment, la caractérisation d'un genre de géminivirus a été généralement accompagné par l'identification de leur vecteur qui dans tous les cas était un insecte (Jeske, 2009) ; pour les begomovirus il s'agit d'un aleurode, *Bemisia tabaci*, pour les mastrevirus, les curtovirus, les turncurtovirus, et les becurtovirus de cicadelles et pour les topocuvirus d'un membracide (Varsani *et al.*, 2014b). Un insecte vecteur est également soupçonné de transmettre EcmLV et ALCV en raison du fait que nous n'ayons pas détecté de structuration spatiale sur les zones géographiques analysées. En effet, ceci peut être le signe d'une transmission efficace par un vecteur très mobile qu'il reste à découvrir. Ainsi, le but de cette dernière étude est de caractériser le mode de transmission des capulavirus et de tester l'hypothèse d'un insecte vecteur.

4.2. Matériel et méthodes

4.2.1. Tests de transmission d'EcmLV via *Cicadulina mbila* (cicadelle) et *Bemisia tabaci* biotype B (aleurode)

Pour commencer, nous avons testé des insectes connus comme vecteurs d'autres géminivirus et qui étaient disponibles au laboratoire. C'est ainsi que nous avons testé l'aleurode *Bemisia tab*aci, vecteur des begomovirus et *Cicadulina mbila*, vecteur du *Maize streak virus*. Nous disposions dans l'unité BGPI de l'élevage des deux espèces cryptiques les plus invasives de *B. tabaci*, MEAM (Middle East-Asia Minor 1) et Med (Mediterranean), anciennement appelé biotype B et Q. La population de *C. mbila* que nous avions en élevage est originaire d'Afrique du Sud. Tous les tests de transmission par insectes ont été effectués sous cage et tous les plants ont été soumis aux conditions de température et d'hygrométrie décrites dans l'article de Bernardo *et al.* (2013).

Quatre plants de tomate (cv. Monalbo) agroinoculés selon la technique décrite dans Bernardo *et al.* (Bernardo *et al.*, 2013) âgés de 2 mois et infectés par EcmLV ont été utilisés en tant que source d'acquisition. Cent individus de l'espèce MEAM de *B. tabaci* ont été mis en acquisition sur les quatre plants infectés pendant trois jours. Au bout des trois jours, les plants soumis à l'acquisition ont été coupés, et 10 plants sains de tomate âgés de 14 jours ont été introduits dans la cage. Trois jours après, les insectes ont été enlevés via un aspirateur à insectes et conservés dans de l'éthanol à 90% à -20°C, les plantes ont quant à elles été traitées avec l'insecticide systémique Confidor à une concentration de 1mL/L.

Quatre-cents cicadelles ont été mises en acquisition durant 6h sur un plant d'*E. caput-medusae* préalablement agroinoculée avec EcmLV (Bernardo *et al.*, 2013). La période d'acquisition a été réduite à 6h pour limiter la mortalité des cicadelles qui ne survivent pas longtemps sur cette euphorbe. Au bout de ces 6h, nous avons transféré 47 de ces cicadelles sur 14 plantes de tomate saine (cv. Monalbo âgées de 14 jours) à raison de 3 à 4 cicadelle par plante. Nous avons aussi transféré des cicadelles sur deux plantes de *E. caput-medusae* saines à raison de 10 cicadelle/plante ; l'âge des plantes saines n'est pas connu car elles ont été produites dans une pépinière commerciale. Au bout de 3 jours toutes les cicadelles étaient mortes. Elles ont été récoltées puis conservées à -20°C dans de l'éthanol à 90%.

Les plantes qui ont été exposées aux insectes supposés virulifères ont été échantillonnées 30 jours après inoculation pour un test de détection de l'EcmLV par PCR (Bernardo *et al.*, 2013)).

4.2.2. Test de transmission mécanique d'EcmLV

Etant donné que certains géminivirus peuvent être transmis mécaniquement, nous avons testé ce mode de transmission pour EcmLV. Des plantes de tabac et de tomate âgées de 2 mois ont été agroinoculées (Bernardo *et al.*, 2013) et testées positives à la présence d'EcmLV. Une feuille symptomatique et jeune de préférence a été récoltée sur ces plantes et broyée dans un mortier stérile en ajoutant une faible quantité de carborandum et 5mL de tampon phosphate (0.1 M Na2H/KH2PO4 buffer pH 7.0). Cet extrait brut d'EcmLV a été appliqué par frottement léger avec les doigts (en portant des gants) sur 3 jeunes feuilles de 10 plantes de tomate (cv. Monalbo) et de 6 plantes de *Nicotiana bentamiana*. Les plantes inoculées ont ensuite été rincées à l'eau distillée et testées 4 semaines après inoculation pour la présence d'EcmLV sur des feuilles nouvelles (Bernardo *et al.*, 2013)).

4.2.3. Recherche *in natura* du vecteur d'EcmLV

Nous avons profité de notre dernière mission d'échantillonnage de plantes dans le fynbos pour récolter des insectes présents sur *E. caput medusae* et qui pouvaient être potentiellement vecteurs. Peu d'insectes ont pu été observés sur *E. caput medusae*. Deux cicadelles et deux pools de 5 pucerons ont été cependant récoltés sur des plantes de la

réserve de Bufflesfontein grâce à des aspirateurs à insectes. Ces insectes ont directement été plongés dans de l'éthanol à 90% et conservés au laboratoire à -20°C.

4.2.4. Recherche *in natura* du vecteur d'ALCV

Pour rechercher le vecteur d'ALCV, nous avons récoltés des insectes piqueurs-suceurs en Camargue en Mai et Juillet 2014. Deux champs de luzerne ont été visités, le premier au niveau de la Tour du Valat et le second sur la commune d'Albaron (Figure 2.7). Les insectes ont été récoltés à l'aide d'un aspirateur à insectes et/ou d'un filet fauchoir. Chaque insecte sélectionné pour le test a été placé individuellement sur un jeune plant de luzerne (var. Eugenia) préalablement repiqué dans un tube en plastique perforé et transporté au champ. Les plantules étaient âgées de 3 semaines au moment de leur exposition. Après dépôt de l'insecte, chaque plante est couverte d'un sachet perforé insectproof et ramené en chambre de culture dans les conditions climatiques fixées pour la préparation des plantules : une hygrométrie de 50% et une phytopériode de 18h jour à 24°C et de 6h nuit à 21°C. Chaque jour, les insectes morts ont été récupérés individuellement et placés dans de l'éthanol à 90%. Treize jours après la récolte au champ, les derniers insectes vivants ont été récoltés individuellement dans de l'éthanol à 90%. Tous ces insectes ont été conservés à -20°C. Les plantules exposées ont été testées par PCR pour la présence d'ALCV en utilisant les amorces Luz-CP-F et Luz-CP-R (Annexe 7).

En Août 2014, des pucerons vivants de l'espèce *Aphis gossipi* ont été récoltés à Assas dans le champ où nous avions détecté des luzernes infectées par ALCV. Soixante plantules de luzerne (var. Magali) âgées de 6 jours ont été mises en présence de ces pucerons. Un mois après exposition de ces plantules, 10 lots de 5 feuilles ont été récoltés et testés pour la présence d'ALCV par PCR (cf. section 3.2.4).

Les insectes de Camargue testés pour leur capacité vectrice ont été observés à la loupe binoculaire et identifiés par des entomologistes du laboratoire.

4.2.5. Extractions d'ADN et détection virale à partir des insectes issus des tests de transmission et récoltés au terrain

Outre les insectes récoltés en Camargue, l'ADN total des insectes a été extrait par la technique décrite par Delatte *et al.* (Delatte *et al.*, 2007).

Concernant les insectes récoltés en Camargue, une extraction d'ADN non destructive a été réalisée afin de pouvoir identifier les insectes *a posteriori* sur la base de leurs caractères morphologiques. Le protocole d'extraction a été mis au point par Isabelle Abt de l'équipe 6 de l'UMR BGPI. Il est en cours de publication et ne peut être détaillé dans ce manuscrit. Brièvement, cette technique a pour but d'extraire les acides nucléiques totaux présents dans les insectes grâce à l'utilisation d'un tampon TNES (Tris, NaCl, EDTA, SDS) et de la protéinase K.

Les ADN extraits ont été utilisés pour y tester la présence du virus recherché selon les techniques PCR décrites précédemment dans ce manuscrit. De plus, les extraits

d'ADN issus des insectes récoltés à Bufflesfontein ont été soumis à un barcoding via le marqueur CO15P dont le protocole est décrit par Germain *et al.* (Germain *et al.*, 2013). Les ADN issus du barcoding ont été séquencés par le Genoscope via la méthode Sanger.

4.2.6. Recherche de la présence d'ALCV dans les graines de luzerne

La transmission par semence n'a jamais été décrite pour un géminivirus mais on ne peut pas exclure qu'une telle transmission soit possible pour des virus d'un nouveau genre. Comme nous ne maîtrisions pas encore la production artificielle de plantes infectées sur lesquelles des semences auraient pu être récoltées et testées, nous avons pu obtenir des semences de luzernes récoltés sur des champs dans lesquels ALCV a été massivement détecté. Pour confirmer que ces semences venaient effectivement de champs infectés, la présence d'ALCV a pu être confirmée car ces semences contenaient des résidus secs de plantes. De plus pour savoir si les semences étaient elles mêmes porteuses d'ALCV, deux tests ont été réalisés, l'un sur les résidus secs de plantes et l'autre sur les semences débarrassées des résidus secs et traité comme suit. Un gramme de ces graines de luzerne de la variété Magali (environ 500 graines) débarassé des résidus secs a été traité deux fois avec du chlore à 1% durant 3min suivi d'un rinçage à l'eau distillée. L'ADN total de ces échantillons, extrait selon le même protocole que celui utilisé pour les plantules (voir précédemment), a été testé par PCR pour la présence de l'ALCV. Enfin, nous avons semé 6000 graines pour tester si ALCV, potentiellement présent sur les plantes qui ont produit ces semences, peut passer sur les plantules qui germeront de ces semences. Cinq semaines après semis, les feuilles produites ont été récoltées et soumises à la méthode de semi-purification de VLPs (Candresse *et al.*, 2014) puis 80µL résultant de ces semi-purifications ont été soumis à une extraction d'ADN (Dellaporta, 1983) et à la détection par PCR d'ALCV.

4.2.7. Séquençage, alignements, phylogénies

Des produits d'amplification de taille attendue pour la détection d'ALCV ont été obtenus à partir d'extraits d'ADNs d'insecte. Ces ADN amplifiés ont été séquencés par la méthode Sanger (Beckman Coulters Genomics). Les séquences ont été alignées avec les séquences décrites dans la partie traitant de la diversité d'ALCV et un arbre phylogénétique a été construit. Les méthodes utilisées sont les mêmes que celles décrites dans la partie précédente.

4.3. Résultats

4.3.1. Recherches sur la transmission d'EcmLV

Toutes les plantes sur lesquelles nous avons tenté une transmission d'EcmLV (*E. caput-medusae*, tabac, tomates) par *Bemisia tabaci*, *Cicadulina mbila* ou par transmission mécanique ont réagi négativement à la détection d'EcmLV. Le test a été réalisé par PCR sur des échantillons de feuilles récoltées un mois après les tests de transmission. De plus, les insectes utilisés pour ces tests de transmission étaient également négatifs au

test de détection d'EcmLV ce qui indique une barrière précoce à la transmission liée à l'acquisition virale.

Concernant les insectes récoltés sur le terrain, nous avons pu identifier les pucerons collectés en Afrique du Sud comme appartenant à l'espèce *Aphis craccivora* et les cicadelles comme appartenant à l'espèce *Platymetopius vitellinus*. Les tests PCR appliqués à ces insectes n'ont pas permis de détecter EcmLV.

4.3.2. Recherche sur la transmission d'ALCV

Les 6000 plantules issues de graines de luzerne de la variété Magali se sont révélées être négatives au test de présence d'ALCV, ce qui nous permet de conclure que ce virus n'est pas transmis efficacement par la graine. Par comparaison aux virus de la famille *Geminiviridae* pour lesquels la transmission par graine n'a jamais été décrite, notre résultat suggère qu'ALCV ne déroge pas à ce dogme.

Lors de la récolte d'insectes en mai 2014, nous avons piégé un total de 401 insectes. Parmi ces insectes, les pucerons étaient majoritaires avec 275 individus détectés. Trente cicadelles ont également été détectées. Dans le détail, nous avons piégé 254 insectes à la tour du Valat (dont 172 pucerons) et 147 à Albaron (dont 103 pucerons). Treize jours après la collecte nous avons pu récupérer et conserver 365 individus (dont 275 pucerons) dans de l'alcool (les 36individus restant étaient trop détériorés pour être conservés). Sur ces 365 individus, 88/275 pucerons se sont révélés être positifs à la présence d'ALCV. Ces individus positifs correspondent à des pucerons provenant de la Tour du Valat (69) et d'Albaron (19). Cela laisse supposer que les pucerons sont capables d'acquérir ALCV. La récolte de juillet 2014, nous a permis de collecter 40 insectes dont 22 qui semblent appartenir au genre de cicadelle *Anaceratagallia* et 8 au groupe des membracides. Aucun de ces insectes n'a été positif au test de détection d'ALCV.

Nous avons pu obtenir des séquences d'ALCV à partir de 9 pucerons « positif » de la première campagne de piégeage. Ces insectes ont été observés sous une loupe binoculaire et leurs caractères morphologiques semblent indiquer qu'il s'agit d'*Acyrtosiphon pisum* (Puceron du pois). Les séquences présentes au sein des pucerons sont uniques et différentes de celles obtenues à partir des luzernes infectées (Figure 2.11).

Les PCR réalisées sur les plantules qui avaient accueilli ces pucerons avant leur détection positive en PCR se sont révélées être négatives à la présence d'ALCV. Par ailleurs toutes les autres plantules qui ont été exposées à des insectes provenant du champ ont réagi négativement à la détection d'ALCV.

Quatre pucerons (*Aphis gossipi*) collectés à Assas en Aout 2014 étaient également négatifs à la détection de l'ALCV. Cependant, 1 mois après le test de transmission réalisés avec ces pucerons provenant du champ, un des 10 lots de feuilles prélevées sur les plantes de luzernes exposées a réagi positivement à la présence d'ALCV. La séquence

correspondante à cet échantillon noté AphidAssas est unique et divergente des autres séquences d'ALCV (Figure 2.11).

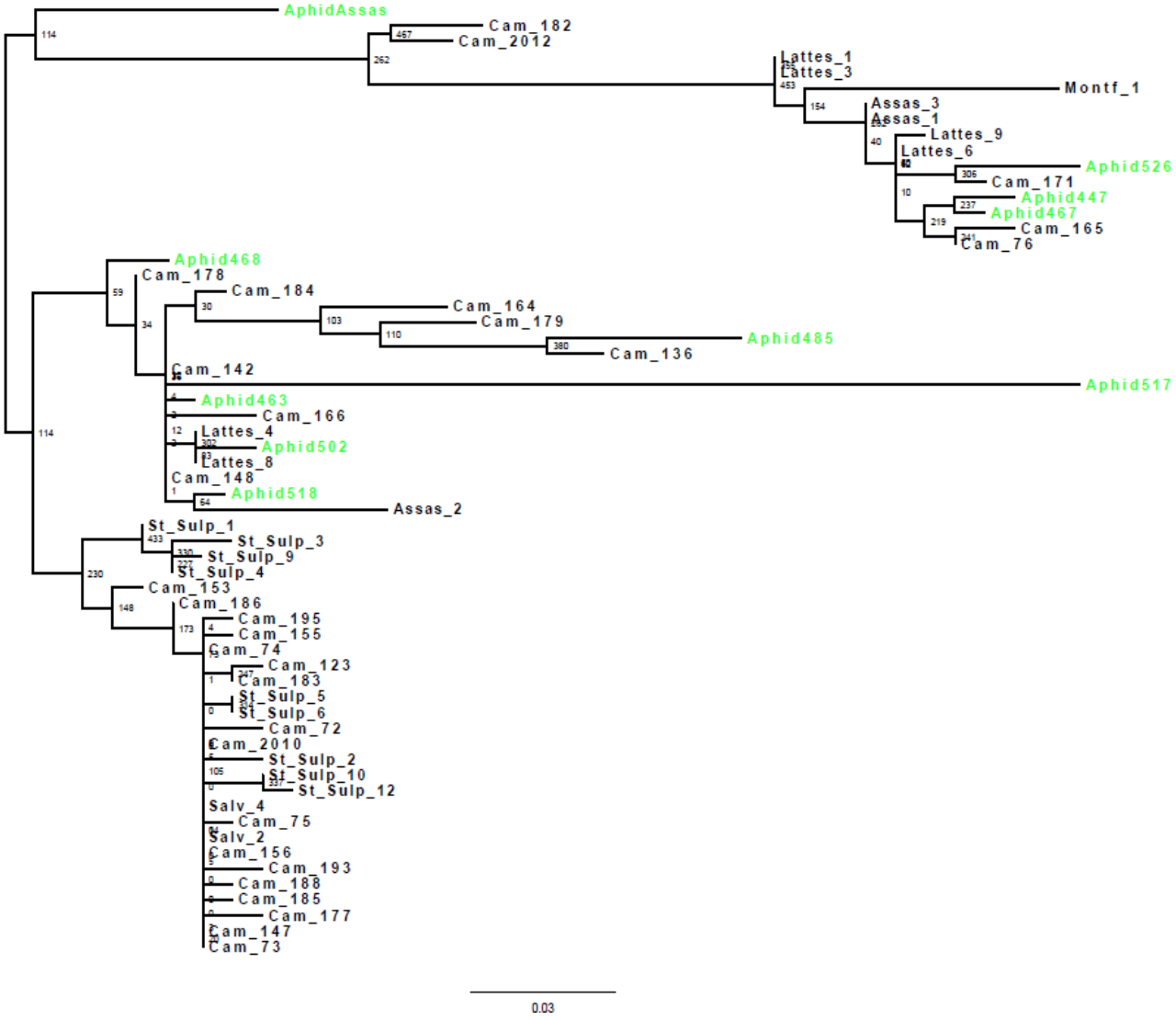


Figure 2.11 : Phylogénie réalisée selon la méthode du maximum de vraisemblance à partir de 51 séquences partielles de la *cp* des isolats d'ALCV provenant de plantes ainsi que de 10 isolats provenant de pucerons (423pb). Cet arbre a été construit sur la base du modèle HKY+I+G avec 500bootstraps. Les isolats provenant de pucerons sont colorés en vert. Les valeurs des bootstraps sont indiquées à la base des nœuds.

4.4. Discussion et perspectives

4.4.1. Recherches sur la transmission d'EcmLV

Comme les capulavirus forment un genre divergent au sein de la famille des *Geminiviridae*, il était difficile d'émettre des hypothèses quant à leur transmission. Cependant, les analyses phylogénétiques ont montré que leur CP était plus proche des Begomovirus que des autres genres. Comme la CP intervient dans la transmission des geminivirus, il était légitime de tester *Bemisia tabaci* en tant qu'insecte vecteur car il est l'unique vecteur des begomovirus. Cependant les tests réalisés avec cet aleurode sur tomate ont été infructueux alors que nous avons utilisé les deux espèces crytpiques les

plus invasives et connues comme faisant partie des meilleurs vecteurs de begomovirus. Les tests de transmission impliquant *Cicaculina mbila*, ont pu être réalisés avec un population venant précisément d'Afrique du Sud, le pays d'isolement d'EcmLV. Même si *C. mbila* n'est probablement pas vecteur d'EcmLV, la transmission a été difficile à tester car cet insecte inféodé aux *Poaceae*, décédait en moins de 24h sur l'euphorbe, et sur la tomate.

Concernant les insectes récoltés à Bufflesfontein, le fait que nous n'y avons pas détecté EcmLV ne donne pas réellement d'information sur le fait qu'ils puissent transmettre EcmLV ou non. En effet, le nombre d'individus dont nous disposions était très restreint et de plus nous ne savions pas si les euphorbes sur lesquelles les insectes ont été récoltés, étaient infectées par EcmLV. Ces conditions font que la probabilité de détecter EcmLV dans ces insectes était vraisemblablement très faible. Par la suite, il faudra envisager de repérer des plants infectés *in natura*, et de récolter prioritairement les insectes qui visitent ces plantes.

L'hypothèse de transmission végétative d'EcmLV n'a pas été testée dans nos travaux et reste donc une possibilité de transmission du virus.

4.4.2. Evaluation de la transmission d'ALCV par la graine

Même si les taux de transmission par la graine peuvent varier de moins de 1% à 100% selon le virus étudié (Hull, 2009), il sont généralement extrêmement faibles. Il est donc nécessaire de tester plusieurs milliers de graines afin de pouvoir détecter si ce type de transmission est possible (Astier, 2007). De plus, à ce jour aucune transmission par la graine n'a pas été démontrée chez les geminivirus. Ici, sur 6000 plantules de luzernes testées, aucune n'était infectée par ALCV, ce qui va dans le sens d'une transmission par un insecte vecteur. En supposant qu'ALCV est transmis par graine à un taux très faible en dessous du seuil que l'on pouvait espérer détecter avec 6000 graines, sa forte prévalence observée en Camargue ne pourrait être expliquée que par la seule transmission par graine. L'hypothèse d'une transmission par vecteur reste donc cohérente.

4.4.3. Une transmission des capulavirus par puceron ?

Nos recherches sur la transmission des capulavirus laissent entrevoir une piste quant à une transmission potentielle par puceron. En effet, nous avons obtenu plusieurs résultats préliminaires qui soutiennent cette hypothèse :

(1) la détection de la présence du virus dans 88/275 des pucerons piégés au terrain

(2) l'obtention de séquences uniques à partir de 9 individus de l'espèce *Acyrtosiphon pisum*

(3) la détection d'ALCV dans des plantules ayant accueilli des individus d'*Aphis gossipi* provenant d'un champ dans lequel nous avions détecté de très nombreux plants infectés par ALCV.

Cependant, la détection d'ALCV dans les individus d'*Acyrtosiphon pisum* et non dans les plantules qui les ont hébergé posent la question de la transmission par cette espèce. En effet, le test de détection ici montre que l'insecte a pu acquérir le virus, toutefois les tests PCR réalisées sur plantules semblent indiquer que le virus n'ait pas été transmis. Il se pourrait que nous n'ayons pu détecter le virus pour plusieurs raisons :

(1) le virus n'a pas été transmis,

(2) le virus n'est pas adapté à la variété testée. En effet, la variété sur laquelle nous avons détecté le virus au champ est la variété Magali, alors que les plantules sur lesquelles nous avons réalisé les tests de transmission via *Acyrtosiphon pisum* font partie de la variété Eugenia. Or, comme les virus ne sont pas tous adaptés de la même manière aux diverses variétés d'une même espèce (Moury, 2010), il se pourrait donc qu'ALCV ne soit pas adapté à la variété Eugenia.

(3) le virus a un temps de réplication très lent. En effet, lorsqu'un virus est inoculé à une plante, il existe une phase de latence durant laquelle le virus ne peut être transmis car la charge virale est trop faible (Anderson and May, 1991). Par exemple, le *Plum pox virus* peut avoir une phase de latence de un à quatre ans dans les vergers (Plumb *et al.*, 1983). L'article que nous avons publié à propos d'EcmLV indiquait qu'un mois après inoculation, une plante sur cinq était infectée par EcmLV. Or, deux ans après l'inoculation, il s'avère que les 5 plantes agroinoculées sont infectées. Le fait qu'*Euphorbia caput-medusae* et que la luzerne soient des plantes pérennes pourraient expliquer le fait que les capulavirus passent par une phase de latence. En effet, on peut supposer qu'un virus touchant une annuelle va devoir se répliquer rapidement s'il veut être transmis avant la mort de sa plante hôte, alors que si un virus touche une plante pérenne, le temps requis pour atteindre une charge virale nécessaire à la transmission est moins contraint.

Les tests réalisés avec *Aphis gossipi* récoltées à Assas semblent indiquer une transmission d'ALCV étant donné que certaines des plantules les ayant accueillis ont été positives au test de détection. La réponse négative d'un échantillon de pucerons provenant de ce champ semble incohérente avec ce succès de transmission. Il faut cependant rappeler que l'échantillon pouvait ne pas être représentatif de la population rapportée du champ car seulement 4 individus ont été testés. Enfin il n'est pas impossible que les plantules ayant accueillis les pucerons du champ aient été positives à la détection d'ALCV du fait de la présence d'un puceron virulifère qui a échappé à notre observation au moment de la récolte.

Le fait que nous retrouvions des séquences uniques à partir des isolats issus d'insectes ou de la transmission par insecte (AphidAssas, Figure 2.11) nous indique premièrement que l'hypothèse d'une contamination lors des collectes ou au laboratoire

est peu probable. Si la transmission d'ALCV par puceron se confirme nous pouvons supposer que les capulavirus soient tous transmis par puceron. En effet on remarque que les géminivirus de chacun des genres pour lesquels un vecteur a été identifié appartiennent à un même groupe d'insecte, aleurode, cicadelle ou membracide (Jeske, 2009; Varsani *et al.*, 2014b). La caractérisation du vecteur d'ALCV nous permettra donc d'aiguiller nos recherches concernant les autres espèces virales du genre Capulavirus.

4.4.4. Recherches en cours sur la vection d'ALCV

Afin de continuer à rechercher l'insecte vecteur d'ALCV, nous avons exposé régulièrement une barquette de 1500 luzernes âgées de 5 à 30 jours dans un champ où nous avons détecté des luzernes infectées par ALCV à Assas. Cette approche, destinée essentiellement à prouver qu'une transmission vectorielle par insecte volant est possible, est totalement sans *a priori*. Chaque barquette est laissée de une à trois semaines au champ puis ramenée au laboratoire en conditions « insect proof ». En prévision d'une transmission sur l'un des lots, les insectes présents sur la barquette en fin d'exposition sont récoltés et conservés dans de l'alcool. Des tests de détection d'ALCV sont réalisés un mois après le dépôt de la barquette au champ afin de détecter si le virus a été transmis. Ainsi nous allons théoriquement pouvoir détecter à quelle période de l'année le virus est transmis afin de mieux cibler son insecte vecteur. De plus, si nous détectons le virus sur les luzernes testées, nous testerons les insectes récoltés à la période correspondante.

La piste selon laquelle *Acyrtosiphon pisum* pourrait transmettre ALCV doit encore être creusée. Pour cela nous disposons d'une plante source du champ qui a été testée positive à la présence d'ALCV et qui a été installée dans une cellule du laboratoire. En outre nous disposons dans l'unité BGPI d'un élevage d'*Acyrtosiphon pisum* et d'un lot de graines de la variété de luzerne sensible, Magali.

4.4.5. Complexité de la recherche du/des vecteur/s des capulavirus

Pour orienter les recherches futures d'un potentiel vecteur des capulavirus il nous faut tenir compte des résultats préliminaires sur les pucerons. C'est un groupe très étudié sur lequel une importante littérature a été produite dans tous les domaines y compris la vection. En effet c'est le premier vecteur de virus de plantes. A cet égard il est serait surprennent qu'aucun des geminivirus n'ait réussi à utiliser un des nombreux insectes de cette grande famille.

D'un point de vue pratique, une manière de se concentrer sur les insectes potentiellement les plus aptes à transmettre ces virus est de cibler les insectes piqueurs suceurs présents dans les champs de luzernes. Pour ce qui est de la zone d'échantillonnage, à la Tour du Valat, une liste des insectes présents dans la réserve naturelle de Camargue est disponible et pourrait servir à ne cibler que les insectes piqueurs suceurs qui se nourrissent sur luzerne. Dans cette optique, *Acyrtosiphon pisum* est un bon candidat car il est réputé pour s'installer sur les plantes de la famille des

Fabaceae et il est notamment connu pour transmettre une trentaine de virus tels que le *Cucumber mosaic virus*. De plus, concernant les *Geminiviridae*, il a déjà été documenté que les interactions entre les virus et leurs vecteurs peuvent être très spécifiques. En effet, on peut citer l'exemple de différentes « races » chez l'insecte *Bemisia tabaci* qui n'ont pas tous la même capacité à transmettre les begomovirus : la race « Sida » polyphage est capable de transmettre une large gamme de begomovirus alors que la race « Jatropha » monophage n'en transmet qu'un seul (De Barro *et al.*, 2011).

Nos études concernant la vection des capulavirus sont pour le moment orientées vers les insectes, or il a été démontré que certains virus peuvent être transmis via le pâturage (Mckirdy *et al.*, 1994) ce qui pourrait élargir notre champs d'étude bien que cela soit difficile à tester.

5. Conclusion générale

Dans ce chapitre nous avons pu caractériser plusieurs espèces d'un nouveau genre de la famille des *Geminiviridae*, à savoir, le genre Capulavirus. En plus d'avoir une organisation génomique propre, les espèces de ce nouveau genre forment un groupe monophylétique divergent des autres genres de géminivirus. Cependant, en dépit d'une hypothèse forte sur la transmission par pucerons, nous n'avons pas encore de résultats définitifs concernant la transmission de ces virus. Ces virus ont été détectés et caractérisés dans diverses régions de l'ancien monde (Afrique, Europe, Asie) à partir d'hôtes cultivés (luzerne et haricot vert) et sauvages (*Euphorbia caput-medusae* et *Plantago lanceolata*). Les prévalences des plants malades sont relativement élevées que ce soit sur luzerne ou sur l'euphorbe sauvage sud-africaine. Comme les autres geminivirus, les capulavirus ont une propension à la recombinaison intraspécifique. Nous avons aussi montré *in vitro* qu'EcmLV est capable d'infecter une plante d'intérêt agronomique, la tomate, et d'y provoquer des symptômes très sévères. Pour estimer le risque réel que représente ces virus pour l'agriculture il faudra approfondir leur caractérisation, et notamment leur prévalence, leur mode de transmission, l'étude de leur gamme d'hôte et le potentiel des ces hôtes à être des réservoirs de virus.

Par ailleurs, la découverte des différents membres de ce nouveau genre à différents endroits du monde et ce en l'espace restreint de 3 ans nous a amené à nous poser la question de leur émergence. Deux hypothèses non exclusives peuvent être émises : (1) ces virus sont en train d'émerger dans l'ancien monde, (2) les nouvelles techniques de biologies moléculaires et l'intérêt nouveau pour le milieu sauvage ont permis de les découvrir alors qu'ils étaient déjà présents depuis un certain temps. L'hypothèse numéro 2 peut être favorisée en ce qui concerne EcmLV, ALCV, et *Plantago Capulavirus* dont la découverte n'a pas été provoquée par la détection de symptômes inhabituels. A l'inverse, la découverte de FbSLCV apporte un soutien à l'hypothèse 1 car il a été isolé sur des plants de haricot qui manifestaient des symptômes sévères en Inde. L'ALCV aurait pu être découvert à partir des symptômes qu'il induit mais il semblerait que les paysans accordent une attention moins soutenue à des plantes fourragères qu'à des plantes produites pour leur semence ou leur fruit. De plus comme ALCV ne semble

pas induire la mort de son hôte, son impact éventuel sur le rendement semble donc difficile à estimer sans études rigoureuses.

EcmLV et *Plantago Capulavirus* sont quant à eux issus du milieu sauvage, un milieu dans lequel peu d'études ont été menées concernant les phytovirus (Cooper and Jones, 2006; Wren *et al.*, 2006). Par ailleurs, ALCV, *Plantago Capulavirus* et EcmLV ont été découvert grâce à la métagénomique qui est apparue en 1998 (Handelsman *et al.*, 1998). On peut alors prévoir qu'un nombre incalculable de nouveaux virus sera détecté dans les années à venir. Ainsi, des virus présents dans leur hôte depuis de nombreuses années mais qui sont jusque-là passés inaperçus, échapperont difficilement aux amplifications aléatoires de la métagénomique qui pourrait être appliqué à des plantes d'intérêt mais aussi à des hôtes et des milieux jusqu'alors inexplorés.

Conclusion générale-Discussion

La synthèse bibliographique proposée en début de manuscrit souligne la difficulté d'avoir une vision globale et complète de l'écologie phytovirale dans les écosystèmes. En effet, alors que les interactions plantes-virus dans les écosystèmes sont multiples et complexes, les études qui les concernent ont été principalement focalisées sur le milieu cultivé. Conscients de cette lacune, nous avons mis en place une nouvelle approche, appelée géo-métagénomique, qui permet d'aborder les dynamiques spatio-temporelles des associations plantes-virus dans les agro-écosystèmes. Contrairement aux travaux classiques de métagénomique environnementale, cette nouvelle approche permet de réassigner chaque séquence virale issue du séquençage haut débit à son hôte géo-localisé à l'échelle d'une grille d'échantillonnage prédéfinie. Dans ce travail de thèse, j'ai entrepris de répondre à trois questions spécifiques : la diversité végétale a-t'elle un impact sur la diversité et la prévalence des phytovirus ? Existe-t'il des patrons de distribution spatio-temporelle des phytovirus dans l'agro-écosystème et quels paramètres écologiques permettent d'expliquer ces distributions ?

1. La géo-métagénomique : avantages et inconvénients

1.1. Une approche encore perfectible

1.1.1. Echantillonnage

Lors de notre échantillonnage nous avons ciblé les plantes majoritaires en chaque point d'échantillonnage, ce qui a eu pour corolaire de sous-estimer le rôle que pouvaient jouer les plantes rares dans l'écologie des phytovirus. Cependant quelques questions liées à ces plantes rares seraient intéressantes à traiter, notamment on pourrait se demander quelles sont les causes de leur rareté : ces plantes sont elles rares parce qu'elles ne sont pas compétitives ? Les virus interviennent-ils dans la compétition (Alexander, 1998) Sont-elles moins touchées par les infections virales en raison d'un effet de dilution (Keesing *et al.*, 2010; Keesing *et al.*, 2006) ? Ou sont-elles rares car elles ont été sujettes à des épidémies virales

La grille d'échantillonnage nous a permis de faire ressortir certains patrons concernant les prévalences et distributions virales. Cependant, on peut se demander si la maille utilisée est adéquate pour répondre à notre problématique et si l'utilisation d'une maille différente aurait mené à des résultats différents. Nous pouvons aussi nous demander quelle est l'influence de l'hétérogénéité du paysage et de la connectivité à l'échelle de chaque point sur les prévalences et distributions virales. Afin de tester cela, nous avons mis en place une approche permettant de prendre en compte les points de la grille qui entourent chaque point étudié. Ainsi, pour chaque point donné, nous avons calculé un score correspondant à la moyenne des scores attribués respectivement aux 8

et 24 points qui l'entourent dans un rayon de 500m et 1000m. En utilisant cette méthode, nous sommes arrivés à la même conclusion, à savoir que les prévalences virales du milieu cultivé sont significativement plus élevées que celles du milieu sauvage. Ainsi cela semble indiquer que le dispositif expérimental mis en place pour réaliser cette étude est adéquat pour répondre à nos questions. On ne peut cependant pas exclure qu'une maille plus courte que 500m aurait pu révéler des structurations spatiales plus petites, de l'ordre de quelques dizaines ou centaines de mètres.

1.1.2. Traitement des échantillons : biologie moléculaire et bioinformatique

Les techniques que nous avons utilisées en biologie moléculaire ou pour les traitements bioinformatique de nos données peuvent également comporter quelques biais méthodologiques. Par exemple, il est fortement possible que certains virus aient échappé à certaines étapes de la semi-purification (filtration, centrifugation) (Blomstrom, 2011) ou encore à celle des amplifications (Hamady and Knight, 2009; Roossinck *et al.*, 2010; Yilmaz *et al.*, 2010). Une autre critique menée sur les études de métagénomique est celle de la présence non négligeable de contaminations (Degnan and Ochman, 2012; Kunin *et al.*, 2008; Rosseel *et al.*, 2014). Ainsi, nous avons tenté d'épurer nos jeux de données en enlevant les séquences présentant des taux de similarité élevé (97%) avec les virus couramment étudiés dans notre laboratoire. Cependant certaines contaminations issues de virus ne provenant pas du laboratoire mais des échantillons eux-mêmes peuvent potentiellement biaiser nos résultats. Ces contaminations sont difficiles à identifier et à écarter de l'analyse finale sans travaux complémentaires de ré-amplification et reséquençage. Au niveau de l'analyse bioinformatique, il est important de souligner que notre analyse Blast a été réalisée contre une base de données ne contenant que des séquences de virus. Une analyse menée en parallèle via le pipeline GALAXY (Goecks *et al.*, 2010) à partir duquel nous avons réalisé des Blasts contre la GenBank entière (tous types de séquences : virales et non virales) nous a permis d'obtenir des résultats très légèrement différents. A titre d'exemple, pour F2010 le pipeline QIIME aurait assigné 2 fois plus de reads aux *Totiviridae* que le pipeline GALAXY. Nous pouvons donc supposer que certaines séquences ont été attribuées à des virus alors qu'elles proviennent en réalité d'autres organismes. Par ailleurs, les bases de données regorgent de séquences contenant des erreurs ce qui peut mener à des résultats de Blasts biaisés (Krawetz, 1989; Wesche *et al.*, 2004). Ainsi, la représentation des communautés virales que nous avons obtenue doit être considérée comme une estimation et non comme une représentation exacte de la réalité. On peut cependant estimer que beaucoup de ces biais sont inhérents aux protocoles et méthodes d'analyses d'un projet de métagénomique. Ainsi, si les même biais sont présents sur les analyses en chaque point, les points restent comparables

1.2. Des avancées notables

Malgré la liste non exhaustive des biais potentiels liés à notre méthodologie évoqués ci-dessus, cette thèse nous a montré le potentiel de la géo-métagénomique concernant deux grands axes que nous avons abordés dans un cadre spatio-temporel : (1) l'étude des prévalences virales de communauté, (2) la découverte de virus. Cependant, l'utilisation de la géo-métagénomique ne nous a pas permis d'avoir des résultats entièrement satisfaisants quant à l'étude de la diversité phytovirale.

1.2.1. Prévalences virales de commuanuté

1.2.1.1. Prévalence virale de communauté à l'échelle de l'agro-écosystème

Les résultats que nous avons obtenus concernant le nombre d'échantillons de plantes virosées diffèrent de ceux obtenus dans d'autres études (Roossinck, 2012b; Roossinck *et al.*, 2010). En effet nous avons détecté la présence de virus sur 26 à 58% des échantillons de plantes étudiées alors que l'étude de Marylin Roossinck indique 70% d'échantillons de plantes virosées (Roossinck, 2012b; Roossinck *et al.*, 2010). Cet écart peut être expliqué par la différence entre les méthodes utilisées ou encore entre les écosystèmes étudiés.

1.2.1.2. Prévalences virales de communauté et milieux cultivés et non-cultivés

Nous avons montré à l'échelle d'un agro-écosystème que la prévalence virale de communauté dans le milieu cultivé est généralement plus élevée que celle du milieu non-cultivé. Ce résultat apporte une généralisation à des études antérieures qui s'étaient limitées à une seule espèce virale (Pagan *et al.*, 2012; Rodelo-Urrego *et al.*, 2013). Ainsi, grâce à notre étude menée sur l'ensemble des virus d'un agro-écosystème, nous pouvons conclure que le milieu cultivé est plus souvent soumis à des épidémies que le milieu non-cultivé. Plusieurs éléments pourraient expliquer cette différence de prévalence virale :

- *La fragilité des plantes cultivées face aux infections*. Dans la synthèse bibliographique nous avons décrit la faiblesse des cultivars à haut rendement sélectionnés par l'Homme. Selon la théorie de l'allocation des ressources, une plante qui investit ses ressources dans la croissance, diminuerait son investissement dans la résistance. De plus, quand une variété est réputée pour ses rendements exceptionnels, elle est généralement choisie par une majorité de cultivateurs qui la produisent sur des surfaces de plus en plus grandes. La réduction de la diversité des cultures jusqu'à un point extrême où un seul génotype couvre des milliers d'hectares, ce qui augmente dangereusement le risque de catastrophe épidémique (Gallet *et al.*, 2013; Mundt, 2002; Power, 1991). A l'inverse, l'effet de dilution des génotypes, espèces, genres et famille de plantes dans un environnement naturel peut expliquer les prévalences virales plus faibles dans le milieu non-cultivé. Selon cette même logique, il a été montré que les

mélanges variétaux dans les cultures peuvent permettre de réduire l'incidence des pathogènes associés (bactéries, champignons ou virus) (Mundt, 2002). Toutefois, peu d'expérimentations ont été menées concernant les phytovirus (Mundt, 2002). Une étude a cependant permis de démontrer que l'utilisation de mélanges variétaux d'avoine (*Avena sativa*) sensibles et résistants permettait de réduire l'incidence des B/CYDV (Power, 1991). On compare souvent les interactions hôte-pathogène à une course à l'armement. Par la sélection de variétés résistantes, l'humain participe significativement à cette course à l'armement en milieu cultivé. Dans la plupart des cas il équipe un génotype unique avec un armement efficace et spécifique contre la menace d'un pathogène qui subit une pression de sélection très forte pour des souches hautement virulentes. Quand ces souches virulentes apparaissent, les prévalences virales augmentent dramatiquement dans le milieu cultivé. A l'inverse, en milieu non-cultivé qui par définition n'est pas touché par les interventions humaines, la co-évolution sur le long terme entre les plantes et leurs virus conduisent à une plus grande diversité des allèles de résistance et d'avirulence. La probabilité d'interactions compatibles étant plus faible, les prévalences virales sont théoriquement réduites. On peut également prévoir que l'évolution d'un virus vers un optimum de fitness, surtout dans le milieu non-cultivé et donc non perturbé par une intervention humaine, ne conduit pas nécessairement à une virulence accrue (Anderson and May, 1982; Ewald, 1983; Frank, 1996) ; ne pas tuer son hôte permet notamment une durée plus longue de l'infection. Des études ont par ailleurs supposé que le milieu sauvage regorgeait d'interactions plante-virus mutualistes (Roossinck, 2005; Roossinck, 2011c; Wren *et al.*, 2006). Marylin Roossinck fait par exemple l'hypothèse que les *Partitiviridae*, seraient pour partie des virus mutualistes, et tout comme elle, nous avons pu voir que ces virus étaient majoritairement présents dans le milieu sauvage. Il reste cependant à approfondir la nature de ces interactions. De surcroît, deux ans après leur inoculation avec EcmLV, les plants d'*Euphorbia caput-medusae* sont toujours infectés et ne présentent aucun symptôme apparent ce qui nous amène à nous questionner sur l'éventualité d'une interaction mutualiste entre EcmLV et son hôte.

- La connectivité entre les différentes populations hôtes en milieu sauvage. Une étude récente concernant l'interaction *Plantago lanceolata*/*Podosphaera plantaginis* (champignon phytopathogène) à l'échelle de la métapopulation a permis de démontrer qu'une connectivité accrue entre les populations d'hôtes sauvages leur permet d'être plus résistantes et engendre une extinction des populations de parasites associés. Ce phénomène pourrait être expliqué de deux manières : (1) une connectivité accrue engendrerait un meilleur flux de gènes de résistance, (2) une grande densité des réseaux entre les populations hôtes est le reflet d'un environnement riche en ressources, ce type d'environnement est alors propice à l'allocation des ressources dans la résistance (Jousimo *et al.*, 2014). Le même phénomène pourrait être à l'origine des faibles prévalences virales observées en milieu non-cultivé comparé au milieu cultivé dans les deux écosystèmes que nous avons étudiés.

1.2.1.3. Prévalences virales de communauté et plantes exotiques et indigènes

Nous avons également montré que les prévalences virales de communauté les plus fortes peuvent être associées à certaines familles virales mais également à des paramètres écologiques tels que le statut de la plante hôte. En effet, pour les deux campagnes d'échantillonnage en Afrique du Sud, les échantillons de plantes indigènes ont présenté des prévalences significativement moins élevées que les échantillons de plantes exotiques. L'accumulation d'agents pathogènes dans des plantes introduites est un phénomène connu (Flory and Clay, 2013). Les plantes exotiques n'ayant pas coévolué avec les agents pathogènes présents dans leur aire d'introduction fait qu'elles sont d'avantage sujettes aux infections que les plantes natives. Toutefois, peu d'études ont été entreprises concernant les effets écologiques de l'accumulation de pathogènes dans les plantes exotiques (Flory and Clay, 2013). Deux hypothèses antagonistes ont été émises à ce sujet mais n'ont été validées que dans un nombre restreint d'études. La première hypothèse, intitulée « pathogen accumulation and invasive decline », suggère que l'accumulation d'agents pathogènes dans les populations de plantes va conduire à leur déclin, limitant ainsi leur potentiel invasif (Flory and Clay, 2013). La deuxième hypothèse intitulée « pathogen spillover » prédit que les plantes exotiques peuvent être un lieu d'accumulation d'agents pathogènes qui vont ensuite infecter les plantes natives. Selon cette théorie, la capacité compétitive des plantes natives étant réduite, les plantes exotiques envahiraient l'écosystème {Kelly, 2009 #1550;Mangla, 2008 #1552;Malmstrom, 2005 #551;Malmstrom, 2005 #1293;Malmstrom, 2006 #634;Malmstrom, 2007 #554}. Comme les deux hypothèses conduisent à des conséquences diamétralement opposées, il est difficile de savoir si le phénomène que nous avons détecté ici, à savoir que les prévalences virales sont plus élevées dans les plantes exotiques du fynbos, va engendrer un déclin de leurs populations ou au contraire si cela va augmenter leur potentiel invasif. On peut cependant noter que dans les cas des graminées exotiques observées dans le fynbos (*Briza maxima*, *Avena fatua*, *Bromus diandrus* etc.), des études décrivent leur succès en terme de croîssance aisée, de propagation efficace et de survie en conditions de stress (Milton, 2004). Trois raisons ont permis d'expliquer ce bon niveau de compétitivité : (i) les graines sont dispersées efficacement par les herbivores via le transport sur leur pelage ou via leur fèces, (ii) les plantes présentent de bonnes tolérances à la sécheresse et (iii) elles ont des taux de repousse plus élevés que ceux des plantes natives suite au pâturage intensif ou à des incendies (Milton, 2004). Ces résultats nous amène à supposer que ces graminées exotiques ne sont manifestement pas éliminées par les virus indigènes mais au contraire sont peut être des acteurs centraux du phénomène de « Spill over » induisant des phénomènes de transmission de virus au sein des communautés de plantes indigènes.

1.2.2. Découverte de nouvelles espèces phytovirales

La géo-métagénomique a permis la détection potentielle de centaines de nouvelles espèces virales qu'il faudrait caractériser par des méthodes plus classiques de

biologie moléculaire. L'importance de ces découvertes dans notre étude en particulier peut être illustrée par la détection de nouveaux gemycircularvirus et mycovirus ainsi que de plusieurs espèces de Capulavirus que nous avons caractérisés plus finement. La caractérisation plus fine permet de présicer la position taxonomique des virus et d'estimer leur possible impact négatif sur les cultures et peut être même la possibilité d'un impact positif en cas de découverte d'interaction mutualiste. C'est ce que nous avons commencé à faire pour les capulavirus qui ne représentent que 2,4% des nouvelles espèces potentiellement découvertes par nos travaux. Ainsi, comme ces caractérisations plus fines sont chronophages et ne pourront pas être réalisées pour tous les virus au sein d'un seul laboratoire, il faut espérer que les ressources biologiques mises en lumière par la géo-métagénomique puissent être accessibles à la communauté scientifique internationale pour en exploiter tout leur potentiel. C'est ce que nous avons initié en partageant nos données avec plusieurs laboratoires français (Dr. M. Ogliastro, INRA et Dr. E. Hébrard, IRD) et sud-africain (Dr. G. Pietersen, University of Pretoria).

Il est apparemment surprenant qu'un genre viral appartenant à une famille qui est scrutée depuis une trentaine d'année pas une communauté très active de virologues à travers le monde soit resté inconnu aussi longtemps. Pour la plupart d'entre eux leur découverte récente tient au progrès des techniques de recherches sans *a priori* des virus et l'exploration récente du milieu sauvage. L'hypothèse d'une émergence récente ne semble s'appliquer qu'au capulavirus d'Inde, détecté par une approche classique ciblée sur des plantes cultivées exprimant des symptômes sévères. Un saut d'hôte à partir du milieu non cultivée est peut-être à l'origine de cette émergence. Une analyse de métagénomique telle que réalisée dans cette étude à l'interface entre le milieu cultivée et non cultivé permettrait de confirmer cette hypothèse dans la zone où a été trouvé ce virus en Inde. Une des caractéristiques incontournables pour approfondir la connaissance de ce nouveau genre est son mode de transmission.

La caractérisation de ces nouvelles espèces virales dans une famille déjà très étudiée, souligne encore davantage combien la diversité des phytovirus a été sous-estimée. L'exploration de la face cachée de la virosphère est un enjeu majeur pour la compréhension de l'évolution virale et des émergences.

1.2.3. Diversité phytovirale

Dans un premier temps nous avons pu montrer que la quasi-totalité des familles de phytovirus décrites par l'ICTV sont présentes aussi bien dans les milieux non-cultivés que dans les milieux cultivés. Malheureusement, nous avons du nous contenter d'une identification virale au rang taxonomique de la famille virale, ce qui a fortement limité la recherche d'une relation entre diversité virale et diversité végétale. Le manque de couverture du pyroséquençage 454 est en partie responsable de ce problème. Les performances accrues du séquenceur Illumina MiSeq devraient désormais minimiser ce problème car il permet d'obtenir des reads atteignant les 2 X 300 nucléotides avec une plus grande profondeur de séquençage. En effet, alors que la méthode 454 ne fournit qu'environ 1 million de séquences de 250-400 pb par run, l'Illumina MiSeq en produit

25 millions par run (Liu *et al.*, 2012b) (www.illumina.com). L'obtention de génomes complets permettra non seulement d'identifier les virus au niveau de l'espèce mais aussi d'estimer leur gamme d'hôte. Ces deux éléments constituent un progrès considérable dans l'exploitation de données de métagénomique. La gamme d'hôte est en effet un élément important dans l'analyse de l'influence de la diversité végétale sur la diversité et la prévalence virale (Keesing *et al.*, 2010; Keesing *et al.*, 2006). De plus, l'obtention de génomes entiers avec des couvertures importantes rendrait possible l'évaluation des potentiels évolutifs des virus détectés via l'étude des mutations, recombinaisons, et réassortiments (Beerenwinkel *et al.*, 2012; Beerenwinkel and Zagordi, 2011; Blinkova *et al.*, 2009; Bull *et al.*, 2012; Wright *et al.*, 2011). Il est aussi à espérer qu'au delà du gain d'une meilleure couverture et d'une amélioration de la profondeur de séquençage, des progrès seront réalisés au niveau de la quantification des abondances virales. Les études menées jusqu'alors ont montré que l'estimation des abondances virales à partir du nombre de reads était périlleuse. En effet, le séquençage 454 semble surestimer l'abondance en taxon et en gènes de 11% à 35% (Gomez-Alvarez *et al.*, 2009). Par ailleurs, certaines séquences sont préférentiellement amplifiées et par conséquent l'abondance des pathogènes demeure faiblement corrélée avec leur concentration dans l'échantillon (Dong *et al.*, 2011; Hamady and Knight, 2009; Yang *et al.*, 2011; Yilmaz *et al.*, 2010).

2. Perspectives

2.1. Tester de nouveaux paramètres écologiques pour leur influence potentielle sur les dynamiques phytovirales

Une fois améliorée, la géo-métagénomique pourrait permettre de tester l'influence de plusieurs facteurs écologiques sur la distribution spatio-temporelle des phytovirus. Par exemple, quelques études révèlent une influence possible du pâturage sur certains phytovirus (Borer *et al.*, 2009; Malmstrom *et al.*, 2006; Mckirdy *et al.*, 1994). Une étude de métagénomique phytovirale comparant des communautés pâturées et non pâturées pourrait confirmer que ce paramètre a une influence sur la diversité et la prévalence des phytovirus. Dans la même logique, on pourrait tester si le feu a une influence sur les prévalences et diversités phytovirales. Cette étude pourrait être réalisée précisément dans le fynbos qui consiste en une végétation soumise à des incendies volontaires (et parfois involontaires) contribuant à sa régénération et à son maintien (Kraaij, 2010). Il serait alors intéressant de comparer les communautés virales issues de communautés végétales ayant subi des incendies à des fréquences plus ou moins longues (en moyenne une quinzaine d'années) afin d'évaluer l'effet potentiel de cette action humaine sur les communautés phytovirales.

2.2. Utilisation de la métagénomique en diagnostic et en épidémio-surveillance

Une des questions actuelles liée aux travaux de métagénomique concerne son utilisation potentielle en diagnostic et en épidémio-surveillance (MacDiarmid *et al.*, 2013). La métagénomique nous a permis de détecter des phytovirus sur un nombre important d'échantillons. Même si la technique n'est pas exhaustive, on peut raisonnablement considérer qu'elle détecte la grande majorité des virus connus ou inconnus pouvant potentiellement menacer les cultures. La détection du BYDV en Camargue, un endroit où il n'a jamais été décrit, illustre l'utilisation potentielle de cette technique pour l'épidémio-surveillance. La découverte de l'ALCV sur une plante fourragère très commune en France en est une deuxième illustration, fort éloquente, car dans ce cas le virus était totalement inconnu. La géo-métagénomique peut également venir en aide à l'épidémio-surveillance en ce qui concerne la recherche des réservoirs naturels de ces virus et indirectement à la connaissance de leur gamme d'hôte. Par exemple, ALCV a été seulement détecté sur des plants de luzerne ce qui nous laisse supposer que cette plante est l'hôte unique de ce virus ou alors que sa gamme d'hôte est extrêmement restreinte. Quant au BYDV, un virus connu pour sa large gamme d'hôte au sein des *Poaceae* nous l'avons détecté, outre sur du blé, sur diverses adventices telles que *Hordeum marinum* et *Puccinellia festuciformis,* ce qui nous amène à supposer que ces adventices peuvent jouer un rôle de réservoir naturel dans l'écosystème étudié. Au cours de ma thèse nous avons découvert une autre application de la géo-métagénomique qui pourrait être intitulé, « Attention, un train peut en cacher un autre ! » (publication en annexe 2). De façon inattendue, nous avons pu montrer que certaines plantes que l'on savait infectées par un virus, étaient en plus infectées par un deuxième virus totalement inconnu.

Fort de notre expérience nous considérons que la métagénomique peut être très utile pour le diagnostic, et la surveillance épidémiologique. En outre, les résultats de métagénomique pourraient être exploités de façon très concrète pour le contrôle des maladies virales. Par exemple, concernant le BYDV, détecté pour la première fois en Camargue, nous pourrions préciser quels sont parmi les hôtes sauvages identifiés ceux qui constituent les réservoirs majeurs et mesurer l'impact potentiel du virus sur les rendements du blé. Sur la base de ces résultats, des mesures préventives pourraient être préconisées.

2.3. Métagénomique sur les insectes vecteurs associés à l'agro-écosystème

Les insectes vecteurs sont des acteurs principaux des dynamiques virales. Comme évoqué dans la synthèse bibliographique, il y a un réel enjeu à mener ces études de métagénomique sur des insectes potentiellement vecteurs (Ng *et al.*, 2011a). Par manque de temps, cette approche n'a pas pu être menée dans le cadre de cette thèse. Elle aurait pu compléter l'approche « plante » en ciblant les insectes présents sur chaque

point d'échantillonnage. Cependant, contrairement aux plantes qui sont immobiles et restent en un point tout au moins sur une saison entière, la récolte d'un échantillon d'insectes en chaque point aurait été fort difficile. En effet, les insectes étant mobiles, et ayant des cycles de développement qui ne se chevauchent pas forcément, la récolte des insectes aurait été fortement dépendante de l'unique date de récolte réalisée chaque année. Si une telle approche devait être envisagée, le protocole d'échantillonnage devra être réfléchi en fonction de ces contraintes pour que les échantillons soient représentatifs des vecteurs qui visitent la zone échantillonnée. Ce type d'étude permettrait d'avoir une image plus précise de la dynamique des phytovirus au niveau de l'agro-écosystème et répondre aux questions : « Quel insecte est le vecteur de quel virus ? À quel endroit ? Au niveau de quelle plante ? ». Cependant, le traitement de ces données nécessite un minimum de précaution, car la présence d'un phytovirus dans un insecte ne signifie pas nécessairement qu'il en est le vecteur. C'est ainsi que des mastrevirus ont été détecté dans une libellule qui pourtant n'est pas vecteur de phytovirus (Rosario *et al.*, 2013).

2.4. Notre jeu de données : une ressource inépuisable de découvertes

L'analyse des reads de nos jeux de données via le pipeline GALAXY nous a montré qu'un nombre important des séquences obtenues correspondent à des virus n'infectant pas les plantes. Il pourrait s'agir de virus particulièrement stables, capables de se conserver temporairement dans l'environnement et plus particulièrement sur les plantes (Roossinck, 2012b). De plus, il est fort probable que certaines des plantes échantillonnées étaient porteuses de pontes d'insectes, de déjections animales… Cette hypothèse a été confirmée par la découverte et la description d'une nouvelle espèce de densovirus (virus d'insectes) qui a d'ailleurs permis d'étoffer leur phylogénie (publication que nous avons soumise en octobre 2014 à la revue Genome Announcement, Annexe 9). En outre, un grand nombre de séquences issues de nos travaux restent encore à identifier. Pour le moment ces séquences se rajoutent à ce qui est communément appelé en métagénomique « la matière noire ».A-t-on affaire à des phytovirus ? D'autres virus ? Il est alors difficile d'y répondre. Des méthodes basées sur les signatures des métagénomes ont été récemment proposées pour tenter de d'identifier cette matière noire (Dinsdale *et al.*, 2008; Willner *et al.*, 2009a). Il reste donc un défi à relever pour donner du sens à cette matière noire de nos jeux de données. Sa présence n'est certainement pas anodine et pourrait éventuellement révéler des phénomènes encore inconnus dans le domaine de la virologie végétale, voire de la virologie générale.

Annexe 1 :

Tableau récapitulatif du nombre de phytovirus répertoriés par l'ICTV (mise à jour du 30 Juin 2014) et de leur taxonomie. (a : contient des espèces infectant les animaux ; b : contient des espèces infectant les champignons.)

Génome	Ordre	Famille	Genre	Nombre d'espèces infectant les plantes
ssRNA-	Mononegavirales	*Rhabdoviridae* [a]	Cytorhabdovirus	9
			Nucleorhabdovirus	10
	Non assignés	*Bunyaviridae* [a]	Tospovirus	9
		Ophioviridae	Ophiovirus	6
		Non assignés	Emaravirus	4
			Tenuivirus	7
			Varicosavirus	1
ssRNA+	Picornavirales	*Secoviridae*	Cheravirus	4
			Comovirus	15
			Fabavirus	5
			Nepovirus	36
			Sadwavirus	1
			Sequivirus	3
			Torradovirus	2
			Waikavirus	3
			Espèces non assignées	3
		Non assignés	Bacillarnavirus	3
			Labyrnavirus	1
	Tymovirales	*Alphaflexiviridae* [b]	Allexivirus	8
			Lolavirus	1
			Mandarivirus	2
			Potexvirus	37
			Espèces non assignées	1
		Betaflexiviridae [b]	Capillovirus	2
			Carlavirus	52
			Citrivirus	1
			Foveavirus	6
			Tepovirus	1
			Trichovirus	7
			Vitivirus	9
			Espèces non assignées	9
		Tymoviridae	Maculavirus	1
			Marafivirus	7
			Tymovirus	27
			Espèces non assignées	1
	Non assignés	*Benyviridae*	Benyvirus	4
		Bromoviridae	Alfamovirus	1
			Anulavirus	2
			Bromovirus	6
			Cucumovirus	4
			Ilarvirus	19
			Oleavirus	1

Génome	Ordre	Famille	Genre	Nombre d'espèces infectant les plantes
ssRNA+	Non assignés	*Closteroviridae*	Ampelovirus	8
			Closterovirus	11
			Crinivirus	13
			Velarivirus	3
			Espèces non assignées	4
		Luteoviridae	Enamovirus	1
			Luteovirus	8
			Polerovirus	17
			Espèces non assignées	7
		Pseudoviridae[b]	Hemivirus[b]	2
			Pseudovirus[b]	16
			Sirevirus	5
			Espèce non assignée	1
		Potyviridae	Brambyvirus	1
			Bymovirus	6
			Ipomovirus	6
			Macluravirus	6
			Poacevirus	2
			Potyvirus	146
			Rymovirus	3
			Tritimovirus	5
			Espèces non assignées	2
		Tombusviridae	Alphanecrovirus	3
			Aureusvirus	4
			Avenavirus	1
			Betanecrovirus	3
			Carmovirus	19
			Dianthovirus	3
			Gallantivirus	1
			Macanavirus	1
			Machlomovirus	1
			Panicovirus	2
			Tombusvirus	17
			Zeavirus	1
			Espèces non assignées	2
		Virgaviridae	Furovirus	6
			Hordeivirus	4
			Pecluvirus	2
			Pomovirus	4
			Tobamovirus	33
			Tobravirus	3

Génome	Ordre	Famille	Genre	Nombre d'espèces infectant les plantes
ssRNA+	Non assignés	Non assignés	Cilevirus	1
			Idaeovirus	1
			Ourmiavirus	3
			Polemovirus	1
			Sobemovirus	14
			Umbravirus	7
dsRNA	Non assignés	*Amalgaviridae*	Amalgavirus	4
		Endornaviridae [b]	Endornavirus [b]	6
		Partitiviridae [ab]	Alphapartitivirus [b]	6
			Betapartitivirus [b]	7
			Deltapartitivirus [b]	5
			Espèces non assignées	13
		Reoviridae [a]	Phytoreovirus	3
			Fijivirus	8
			Oryzavirus	2
		Non assigné	Higrevirus	1
ssDNA	Non assignés	*Geminiviridae*	Becurtovirus	2
			Begomovirus	288
			Curtovirus	3
			Eragrovirus	1
			Mastrevirus	29
			Topocuvirus	1
			Turncurtovirus	1
		Nanoviridae	Babuvirus	3
			Nanovirus	6
			Espèces non assignées	1
		Non assigné	Bacilladnavirus	1
dsDNA	Non assignés	*Caulimoviridae*	Badnavirus	25
			Caulimovirus	9
			Cavemovirus	2
			Petuvirus	1
			Solendovirus	2
			Soymovirus	4
			Tungrovirus	1
		Phycodnaviridae	Chlorovirus	19
			Coccolithovirus	1
			Phaeovirus	9
			Prasinovirus	2
			Prymnesiovirus	1
			Raphidovirus	1
		Non assigné	Dinodnavirus	1

Annexe 2 :

Article : Appearances can be deceptive : Revealing a hidden viral infection with deep sequencing in plant quarantine context

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.H., Fernandez, E., Martin, D.P., Varsani, A., Roumagnac, P., 2014. PloS one 9(7), e102945.

OPEN ACCESS Freely available online

PLOS ONE

Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context

Thierry Candresse[1,2], Denis Filloux[3], Brejnev Muhire[4], Charlotte Julian[3], Serge Galzi[3], Guillaume Fort[3], Pauline Bernardo[3], Jean-Heindrich Daugrois[3], Emmanuel Fernandez[3], Darren P. Martin[4], Arvind Varsani[5,6,7], Philippe Roumagnac[3]*

1 INRA, UMR 1332 Biologie du Fruit et Pathologie, CS 20032, 33882 Villenave d'Ornon Cedex, France, **2** Université de Bordeaux, UMR 1332 Biologie du Fruit et Pathologie, CS 20032, 33882 Villenave d'Ornon Cedex, France, **3** CIRAD, UMR BGPI, Campus International de Montferrier-Baillarguet, 34398 Montpellier Cedex-5, France, **4** Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa, **5** School of Biological Sciences and Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand, **6** Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, Florida, United States of America, **7** Electron Microscope Unit, Division of Medical Biochemistry, Department of Clinical Laboratory Sciences, University of Cape Town, Observatory, South Africa

Abstract

Comprehensive inventories of plant viral diversity are essential for effective quarantine and sanitation efforts. The safety of regulated plant material exchanges presently relies heavily on techniques such as PCR or nucleic acid hybridisation, which are only suited to the detection and characterisation of specific, well characterised pathogens. Here, we demonstrate the utility of sequence-independent next generation sequencing (NGS) of both virus-derived small interfering RNAs (siRNAs) and virion-associated nucleic acids (VANA) for the detailed identification and characterisation of viruses infecting two quarantined sugarcane plants. Both plants originated from Egypt and were known to be infected with Sugarcane streak Egypt Virus (SSEV; Genus *Mastrevirus*, Family *Geminiviridae*), but were revealed by the NGS approaches to also be infected by a second highly divergent mastrevirus, here named Sugarcane white streak Virus (SWSV). This novel virus had escaped detection by all routine quarantine detection assays and was found to also be present in sugarcane plants originating from Sudan. Complete SWSV genomes were cloned and sequenced from six plants and all were found to share >91% genome-wide identity. With the exception of two SWSV variants, which potentially express unusually large RepA proteins, the SWSV isolates display genome characteristics very typical to those of all other previously described mastreviruses. An analysis of virus-derived siRNAs for SWSV and SSEV showed them to be strongly influenced by secondary structures within both genomic single stranded DNA and mRNA transcripts. In addition, the distribution of siRNA size frequencies indicates that these mastreviruses are likely subject to both transcriptional and post-transcriptional gene silencing. Our study stresses the potential advantages of NGS-based virus metagenomic screening in a plant quarantine setting and indicates that such techniques could dramatically reduce the numbers of non-intercepted virus pathogens passing through plant quarantine stations.

Citation: Candresse T, Filloux D, Muhire B, Julian C, Galzi S, et al. (2014) Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. PLoS ONE 9(7): e102945. doi:10.1371/journal.pone.0102945

Editor: Hanu Pappu, Washington State University, United States of America

Received January 31, 2014; **Accepted** June 24, 2014; **Published** July 25, 2014

Funding: This work was supported by the ANR French National Research Agency (Netbiome Program, SafePGR Project). AV and DPM are supported by the National Research Foundation of South Africa. BM is funded by the University of Cape Town. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: philippe.roumagnac@cirad.fr

Introduction

When attempting to prevent the spread of plant diseases, comprehensive inventories of viral diversity are fundamental both for effective quarantine and sanitation efforts, and to ensure that plant materials within biological resource centres (BRCs) can be safely distributed [1,2]. Detection of pathogens is one of the most critical quarantine and BRC operations. Ideally, the tools used for this purpose must be both sensitive enough to accurately detect the presence of even extremely low amounts of pathogen nucleic acids or proteins, and provide sufficient specific information to identify the genetic variants/strains of whatever pathogens are present.

The major challenge of using classical nucleic acid sequence-informed detection tools such as polymerase chain reaction (PCR) or Southern hybridisation assays, is that despite being highly sensitive, these techniques are generally either species or, at best, genus-specific. In addition, such tools lack the capacity to detect, let alone identify, pathogens that are unknown, poorly characterized or highly variable. Although it might be argued that the most economically important pathogens tend to be well characterized and that it is therefore not a serious issue that many of the more obscure pathogens go undetected, it is becoming better appreciated that the "importance" of any particular pathogenic microbe is very difficult to define. Specifically, the environmental and

economic impacts of a particular pathogen can vary widely with varying climatic and ecological conditions and there are large numbers of microbes that are presently not classified as pathogens (or at least which are not noticeably pathogenic to humans or to domesticated plants and animals), which will eventually emerge as important future pathogens [3]. Also, since non-domesticated plant and animal species and countless numbers of microbes which contribute to natural terrestrial ecosystems [4–6] can also potentially be threatened by exotic pathogens, the unconstrained global dissemination of apparently harmless fungi, viruses and bacteria could have serious environmental and economic impacts.

Whereas "sequence-dependent" microbial detection methods, which are generally based on PCR or nucleic acid hybridisation, can only be used to target known pathogens, sequence-independent next generation sequencing (NGS) based approaches can potentially provide an ideal platform for identifying almost all known and unknown microbes present in any particular host organism [5,7–9]. Coupled with innovative sample processing procedures, "metagenomics" applications of NGS [10] have already enabled the identification of novel pathogens through the rapid and comprehensive characterization of microbial strains and isolates within environmental and host tissue samples [9,11].

In addition to numerous applications in the study of animal infecting viruses, NGS-based metagenomics approaches have also been used to detect plant infecting viruses [12]. Three main classes of nucleic-acids have been targeted by such analyses: (1) virion-associated nucleic acids (VANA) purified from viral particles [13,14]; (2) double-stranded RNAs (dsRNA) [15]; and (3) virus-derived small interfering RNAs (siRNAs) [16]. Large numbers of both known and new plant and fungus infecting DNA and RNA viruses and viroids have been detected using these approaches [12,17–21].

A major shortcoming of these metagenomic approaches is, however, that they remain technically cumbersome and too expensive for routine diagnostic applications on collections of eukaryotic hosts - even if barcoded primers are used to bulk-sequence pooled samples from multiple sources [15]. Although prohibitive for high throughput diagnostics, the costs of NGS in the context of viral diversity research are often offset by the vast volumes of useful data that can be generated on viral population dynamics, co-infections, mutation frequencies and genetic recombination [22–24].

Here we describe the application of siRNA- and VANA-targeted NGS approaches to the analyses of two Egyptian sugarcane plants maintained for a number of years at the CIRAD Sugarcane Quarantine Station in Montpellier, France. These plants were both known to be infected with *Sugarcane streak Egypt Virus* (SSEV; Family *Geminiviridae*, Genus *Mastrevirus*) and were maintained for use as positive controls during the application of diagnostic tools for SSEV detection in sugarcane plants passing through the quarantine station. Using the siRNA- and VANA-targeted NGS approaches, we discovered and characterized a novel highly divergent mastrevirus from these two plants. This novel virus was also identified in other sugarcane plants originating from Sudan that exhibited white spots on the base of their leaf blades that become fused laterally, so as to appear as chlorotic stripes. Accordingly, we have proposed naming this virus Sugarcane white streak Virus (SWSV). In addition, we present a detailed analysis of siRNAs derived from the SWSV and SSEV variants infecting the two analysed sugarcane plants.

Materials and Methods

Plant material and sugarcane quarantine DNAs collection

Leaves presenting typical symptoms of sugarcane streak disease were sampled from two sugarcane plants that had previously been found to be infected with *Sugarcane streak Egypt virus* (SSEV) and had been kept in a quarantine greenhouse at the CIRAD Sugarcane Quarantine Station, in Montpellier, France. The two sugarcane plants, VARX and USDA (which was initially maintained at USDA-APHIS Plant Germplasm Quarantine before being transferred to CIRAD in 2007), were both initially collected in Egypt during two independent sampling surveys in the late 1990s [25,26]. These sampling surveys were both carried out on experimental stations and commercial lands in close collaboration with Egyptian authorities (Sugar Crop Research Institute (SCRI), Dr Abdel Wahab I. Allam (Director of SCRI) regarding VARX; and Agricultural Genetic Engineering Research Institute, Dr N.A. Abdallah, and Dr M.A. Madkour regarding USDA). In addition, leaves from six sugarcane plants originating from Sudan (B0065, B0067, B0069, D0002, D0003 and D0005, Table S1) and maintained at the CIRAD Sugarcane Quarantine Station were also used (Material Transfer Agreements between CIRAD and Kenana Sugar Co. Ltd). DNAs from these six plants were extracted using the DNeasy Plant Mini Kit (Qiagen). In addition, DNA was extracted from an additional 18 frozen leaf samples (−20°C), including 17 samples originating from Sudanese sugarcane plants, which had passed through the Montpellier Quarantine station between 2000 and 2009 and one which had been obtained from a sugarcane seedling grown from sugarcane true seeds [fuzz] developed in Guadeloupe from a biparental cross involving plants H70-6957 and B86-049 using the DNeasy Plant Mini Kit (Table S1).

VANA extraction from viral particles, cDNA amplification and sequencing

One gram of leaf material from the VARX and USDA plants were ground in Hanks' buffered salt solution (HBSS) (1:10) with four ceramic beads (MP Biomedicals, USA) using a tissue homogeniser (MP biomedicals, USA). The homogenised plant extracts were centrifuged at 3,200×g for 5 min and 6 ml of the supernatants were further centrifuged at 8,228×g for 3 min. The resulting supernatants were then filtered through a 0.45 µm sterile syringe filter. The filtrate was then centrifuged at 148,000×g for 2.5 hrs at 4°C to concentrate viral particles. The resulting pellet was resuspended overnight at 4°C in 200 µl of HBSS. Non-encapsidated nucleic acids were eliminated by adding 15 U of bovine pancreas DNase I (Euromedex) and 1.9 U of bovine pancreas RNase A (Euromedex, France) followed by incubation at 37°C for 90 min. Total nucleic acids were finally extracted from virions using a NucleoSpin 96 Virus Core Kit (Macherey-Nagel, Germany) following the manufacturer's protocol. The amplification of extracted nucleic acids was performed as described by Victoria *et al*. [14] and aimed at detecting both RNA and DNA viruses. Reverse transcriptase priming and amplification of nucleic acids were used for detecting RNA viruses. A Klenow Fragment step was included in the protocol in order to detect DNA viruses as demonstrated by Froussard [27]. Briefly, viral cDNA synthesis was performed by incubation of 10 µl of extracted viral nucleic acids with 100 pmol of primer DoDec (5′-CCT TCG GAT CCT CCN NNN NNN NNN NN-3′) at 85°C for 2 min. The mixture was immediately placed on ice. Subsequently, 10 mM dithiothreitol, 1 mM of each deoxynucloside triphosphate (dNTP), 4 µl of 5× Superscript buffer, and 5 U of SuperScript III (Invitrogen, USA) were added to the mixture (final volume of 20 µl), which was then

incubated at 25°C for 10 min, followed by 42°C incubation for 60 min and 70°C incubation for 5 min before being placed on ice for 2 min. cDNAs were purified using the QiaQuick PCR cleanup kit (Qiagen). Priming and extension was then performed using Large (Klenow) Fragment DNA polymerase (Promega). First, 20 µl of cDNA in the presence of 4.8 µM of primer DoDec were heated to 95°C for 2 min and then cooled to 4°C. 2.5 U of Klenow Fragment, 10X Klenow reaction buffer and 0.4 mM of each dNTP (final volume of 25 µl) were added. The mixture was incubated at 37°C for 60 min followed by 75°C enzyme heat inactivation for 10 min. PCR amplification was carried out using 5 µl of the reaction described above in a 20 µl reaction containing 2 µM primer (LinkerMid50 primer for VARX: 5′-ATC GTA GCA GCC TTC GGA TCC TCC-3′ and LinkerMid52 primer for USDA: 5′-ATG TGT CTA GCC TTC GGA TCC TCC-3′), and 10 µl of HotStarTaq Plus Master Mix Kit (Qiagen). The following cycling conditions were used: one cycle of 95°C for 5 min, five cycles of 95°C for 1 min, 50°C for 1 min, 72°C for 1.5 min, 35 cycles of 95°C for 30 sec, 50°C for 30 sec, 72°C for 1.5 min +2 sec at each cycle. An additional final extension for 10 min at 72°C was then performed. DNA products were pooled (VARX and USDA products and 94 additional products obtained from other quarantine samples), cleaned up using the Wizard SV Gel and PCR Clean-Up System (Promega) and sequenced on 1/8th of a 454 pyrosequencing plate using GS FLX Titanium reagents (Beckman Coulter Cogenics, USA).

siRNA extraction and sequencing

The nucleic acid extraction and sequencing approach of Kreuze *et al*. [16] was used with slight modifications. Total RNAs were extracted from 100mg of VARX fresh leaf material using Trizol (Invitrogen) following the manufacturer's instructions. Small RNA libraries were directly generated from total RNAs. Small RNAs ligated with 3′ and 5′ adapters were reverse transcribed and PCR amplified (30 sec at 98°C; [10 sec at 98°C, 30 sec at 60°C, 15 sec at 72°C] ×13 cycles; 10 min at 72°C) to create cDNA libraries selectively enriched in fragments having adapter molecules at both ends. The last step was an acrylamide gel purification of the 140–150 nt amplified cDNA constructs (corresponding to cDNA inserts from siRNAs +120 nt from the adapters). Small RNA libraries were checked for quality and quantified using a 2100 Bioanalyzer (Agilent). The library was then sequenced on one lane of a HiSeq Illumina as single-end 50 base reads.

Sequence assembly

Analyses of reads produced by either Illumina (siRNA sequencing) or 454 GS FLX Titanium (Amplified-VANA sequencing) were performed using CLC Genomics Workbench 5.15. *De novo* assemblies of contigs were performed with a minimal contig size set at 100 bp and 200 bp for Illumina and 454 GS FLX Titanium reads, respectively. *A posteriori* mapping of reads against the complete genomes of SWSV (once the full genome had been cloned and sequenced) or SSEV or against parts of these genomes were also performed using CLC Genomics Workbench 5.15. Primary sequence outputs have been deposited in the sequence read archive of GenBank (accession numbers: VANA_USDA dataset: SRR1207274; VANA_VARX dataset: SRR1207275; siRNA_VARX dataset: SRR1207277).

SWSV genome amplification, cloning and sequencing

Two partially overlapping SWSV specific PCR primer pairs were designed so as to avoid any potential cross-hybridization to 63 representative species of the family *Geminiviridae,* including SSEV. These two primer pairs (pair1: SWSV_F1 forward primer 5′-GCT GAA ACC TAT GGC AAA GA-3′ and SWSV-R1 reverse primer 5′-AGC CTC TCT ACA TCC TTT GC-3′; and pair2 ECORI-1F forward primer 5′-GAA TTC CCA GAG CGT GGT A-3′ and ECORI-2R reverse primer 5′-GAG TTG AAT TCC GGT ACC AAG GAC-3′) were complementary to sequences within the *rep* gene of SWSV. Total DNAs from the two sugarcane plants described above (VARX and USDA) were extracted using the DNeasy Plant Mini Kit (Qiagen) and screened for SWSV using the two pairs of primers and GoTaq Hot Start Master Mix (Promega) following the manufacturer's protocol. Amplification conditions consisted of an initial denaturation at 95°C for 2 min, 35 cycles at 94°C for 10 sec, 55°C for 30 sec, 68°C for 3 min, and a final extension step at 68°C for 10 min. Amplification products of ~2.8 Kbp were gel purified, ligated to pGEM-T (Promega) and sequenced by standard Sanger sequencing using a primer walking approach.

Reverse transcriptase priming and amplification of nucleic acids were carried out in order to detect the intron of the *rep* gene. Total RNAs from VARX were extracted using the RNeasy Plant Mini Kit (Qiagen). DNase treatment of extracted RNAs was carried out using RQ1 RNase-Free DNase (Promega) following the manufacturer's protocol. Viral cDNA synthesis was performed by incubation of 1 µl of DNase treated RNAs with 15 µl of RNase free water, 0.6 µM of each primers (SWSV_F2: 5′-ACC ATG TGC TGC CAG TAA TT-3′ and ECORI-2R: 5′-GAG TTG AAT TCC GGT ACC AAG GAC-3′), and 0.4 mM of mixed deoxynucloside triphosphate (dNTPs), 5 µl of 5X Qiagen OneStep RT-PCR Buffer and 1 µl of Qiagen OneStep RT-PCR Enzyme Mix. Tubes were first placed at 50°C for 30 min for cDNA synthesis. PCR amplification was then carried out using the following cycling conditions: One cycle of 95°C for 15 min, 35 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 1 min. An additional final extension for 10 min at 72°C was then performed. Amplification products were gel purified, ligated to pGEM-T (Promega) and sequenced by standard Sanger sequencing.

PCR detection tests

DNAs extracted from 17 sugarcane plants originating from Sudan kept at −20°C or six freshly extracted from plants maintained at the CIRAD Sugarcane Quarantine Station were screened for SWSV. DNA extracted from one sugarcane seedling grown from true seeds (fuzz) was also screened for SWSV. PCR amplification was carried out using the two pairs of primers described above (SWSV_F1 and SWSV_R1; ECORI-1F and ECORI-2R) using GoTaq Hot Start Master Mix (Promega) following the manufacturer's protocol. Amplification products of ~2.8 Kbp were gel purified, ligated to pGEM-T (Promega) and sequenced as described above. Plants infected with SWSV were also screened for all known sugarcane-infecting mastreviruses:*Sugarcane streak Egypt Virus*, *Sugarcane streak virus*, *Maize streak virus*, *Sugarcane streak Reunion virus*, *Eragrostis streak virus* and *Saccharum streak virus*. PCR amplification was carried out using 1 µl of DNA template in a 25 µl reaction containing 0.2 µM of each broad spectrum primer (SSV_1732F: 5′-CAR TCV ACR TTR TTY TGC CAG TA-3′ and SSV_2176R: 5′-GAR TAC CTY TCH ATG MTH CAG A-3′) and GoTaq Hot Start Master Mix (Promega) following the manufacturer's protocol. The following cycling conditions were used: One cycle of 95°C for 2 min, 35 cycles of 94°C for 1 min, 53°C for 1 min, 72°C for 1 min. An additional final extension for 10 min at 72°C was then performed.

Sequence analyses

Six complete genomes of the novel mastrevirus were recovered from plants VARX, USDA, A0037, B0069, D0005 and E0144 (Table S1) and were aligned with the genomes of representative mastreviruses using MUSCLE (with default settings) [28]. Similarly, the predicted replication associated protein (Rep) and capsid protein (CP) amino acid sequences encoded by the viruses within the full-genome dataset were also aligned using MUSCLE. Maximum likelihood phylogenetic trees were inferred for the full genomes (TN93+G+I nucleotide substitution model chosen as the best-fit using jModelTest [29]), Rep (WAG+G+F amino acid substitution model chosen as the best-fit using ProtTest [30]) and CP (rtREV+G+F amino acid substitution model chosen as the best-fit using ProtTest) datasets with PHYML [31]. Approximate likelihood ratio tests (aLRT) were used to infer relative supports for branches (with branches having <80% support being collapsed). All pairwise identity analysis of the full genome nucleotide sequences, capsid protein (CP) amino acid sequences, replication associated protein (Rep) amino acid sequences and movement protein (MP) amino acid sequences were carried out using the MUSCLE-based pairwise alignment and identity calculation approach implemented in SDT v1.0 [32]. The full genome sequence alignment of representative mastrevirus genome sequences together with SWSV was used to detect evidence of recombination in SWSV using RDP 4.24 with default settings [33]. Sequences are deposited in GenBank under accession numbers (SWSV-A [SD-VARX-2013] - KJ187746; SWSV-A [SD -USDA-2013] - KJ187745; SWSV-B [SD -B0069-2013] – KJ210622; SWSV-B [SD -D0005-2013] - KJ187747; SWSV-B [SD -E0144-2013] - KJ187748 and SWSV-C [SD -A0037-2013] - KJ187749).

Test for associations between siRNAs and SWSV/SSEV genomic and transcript secondary structures

The SWSV/SSEV full genome sequences and predicted unspliced complementary and virion strand transcripts were separately folded using Nucleic Acid Secondary Structure Predictor [34], with the sequence conformation set as circular DNA, at a temperature of 25°C. NASP generates a list of all secondary structures detectable within given DNA or RNA sequences and through simulations it demarcates a set of structures referred to as a "high confidence structure set" (HCSS), that confers a higher degree of thermodynamic stability (lower free energy) to the sequences than what would be expected to be achievable by randomly generated sequences with the same base composition (with a $p <= 0.05$).

Given the genomic coordinates of pairing nucleotides within the HCSS, we investigated whether there was any significant trend for more reads (looking both at all reads collectively and at the 21 nt, 22 nt, 23 nt and 24 nt long reads separately) occurring within secondary structures predicted to occur within (i) the full genomes, (ii) the virion-strand transcripts and (iii) the complementary-strand transcripts. The reads were mapped to the secondary structures and we counted how many nucleotides were located at paired and unpaired sites. While Kolmogorov-Smirnov tests (implemented in R; www.r-project.org) were used to determine whether the distribution of reads between paired and unpaired sites were different, Wilcoxon rank-sum tests (also implemented in R, www.r-project.org) indicated whether there were significantly more reads at paired sites compared to unpaired sites and *vice versa*. Whereas the Kolmogorov-Smirnov tests were used to indicate whether any associations existed between siRNA locations and base pairing within nucleic acid secondary structures, the Wilcoxon rank-sum tests were used to determine whether detected associations were positive (siRNAs tended to occur at structured sites) or negative (siRNAs tended to occur outside of structured sites).

Results

454-based sequencing of VANA from the VARX and USDA sugarcane samples

This approach was used in an attempt to detect both RNA and DNA viruses that may be present in the two sugarcane plants [27]. A total of 2612 and 1635 reads were respectively obtained from the VARX and USDA plants following length and quality filtering. One hundred and eight and 18 contigs were produced by *de novo* assembly from the VARX- and USDA-derived reads, respectively. Two contigs from the VARX plant (2706 nt and 412 nt) and two from the USDA plant (2706 nt and 649 nt), encoded proteins with between 91 and 100% sequence identity with previously described SSEV proteins (Table1). BLASTx analysis revealed that an additional two contigs from the VARX plant (2122 and 127 nt) and three contigs from the USDA plant (1836, 196 and 312 nt) were homologous with known mastreviruses but were nevertheless only distantly related to mastrevirus sequences currently deposited in GenBank (Table1).

A posteriori mapping of VANA 454 reads obtained from the VARX and USDA plants against the complete SWSV genome (see below), revealed that 23.9% (625/2612) and 16.1% (264/1635) of the total reads were derived from this genome and that these yielded complete genome coverage at an average depth of 81X and 29X, respectively. Interestingly, a ~120 nt long region of very low coverage (<4X) was identified, which mapped to the large intergenic region (LIR) of the SWSV genome (Figure 1).

A mapping analysis performed with the genome of SSEV indicated that the corresponding values were 53.5% of reads (1398/2612, 159X average coverage depth) and 75% of reads (1227/1635, 138X coverage) for the VARX and USDA plants, respectively (Figure S1).

siRNA Illumina sequencing from the VARX sugarcane plant

A total of 15,275,640 raw reads were generated from the VARX sugarcane sample, which were then filtered down to 3,945,108 high quality reads in the 21 to 24 nt size range of siRNAs. From these reads, 226 contigs were obtained by *de novo* assembly, six of which showed significant degrees of similarity to mastreviruses based on BLASTx [35] searches (Table2). Of these six contigs, two (contigs #121 and #176) had a high degree of identity to SSEV while the remaining four were more distantly related to known mastreviruses. Three of these four contigs (contigs #44, #86 and #101) apparently corresponded with a mastrevirus capsid protein (CP) gene and the other one (contigs #79) with a movement protein (MP) gene, while the cumulative contig length of 761 bp corresponded to slightly more than a quarter of a typical mastrevirus genome (Table2).

Following the cloning and sequencing of the full genome of the new mastrevirus (SWSV; see below) it was determined that 0.59% of the Illumina reads obtained from the VARX plant could be mapped to this genome (Figure 1) to generate contigs that covered 96.3% of the genome at an average depth of 185X with only seven gaps of between three and 40 nucleotides. These gaps were located within the large intergenic region (three gaps) and within the probable replication associated protein (Rep) gene (four gaps; Figure 1) encoded by the C1 ORF. It is noteworthy that the ~120 nt long region of very low coverage (<4X) identified using the VANA approach mapped to the same part of the LIR region of the SWSV genome that remained uncovered during the

Table 1. Lengths, numbers of reads and BlastX analysis results for VANA 454 *de novo* contigs from sugarcane plants VARX and USDA with detectable homology to mastreviral sequences.

Sample	Contig	Contig length (bp)	Number of reads	BlastX Virus	BlastXLocus	BlastX e-value	Percent identity
VARX	#1	2706	1387	SSEV (NP_045945)	RepA	0.00	100%
	#2	2122	470	DDSMV (YP_003915158)	CP	3.84E–56	70%
	#3	412	11	SSEV (AAC98076)	MP	2.07E–8	95,2%
	#7	127	1	BCSMV (YP_004089628)	RepA	1.72e–10	71%
USDA	#1	2706	1128	SSEV (NP_04945)	RepA	9.20E–177	99.2%
	#2	1836	82	DDSMV (YP_003915158)	CP	1.04E–56	48.8%
	#3	649	12	SSEV (NP_04945)	RepA	2.80E–66	91.1%
	#4	196	37	MSV (CAA10092)	RepA	1.34E–8	56.2%
	#13	312	83	SSEV (AAF76868)	RepA	1.65E–30	84.3%

Acronyms used are as follows: SSEV (Sugarcane streak Egypt virus), DDSMV (Digitaria didactyla striate mosaic virus), BCSMV (Bromus catharticus striate mosaic virus), MSV (Maize streak virus).
doi:10.1371/journal.pone.0102945.t001

Illumina-based siRNA sequencing (Figure 1). As has been previously observed for other viruses, genome coverage was highly heterogeneous (Figure 1). However, a clear general trend could be observed, with the region corresponding to the virion sense V1 and V2 ORFs (encoding CP and MP proteins, respectively), showing an average coverage depth of ~436X and the complementary sense C1 ORF showing an average coverage of only ~38X. Coverage of the non-coding large and small intergenic regions and the presumed C1 ORF intron were even lower at 17.5X and 6.8X, respectively.

It is also noteworthy that besides differences in coverage depth, these various genomic regions of SWSV also showed differences in the siRNA size classes that they yielded. While there was an enrichment of the 21 and 22 nt siRNA size classes amongst the total siRNA reads mapping to the V1 and V2 ORFs, there was a depletion of the 21 nt siRNA size classes and an enrichment of the 24 nt size class amongst total siRNA reads mapping to the C1 ORF (Figure 2). The LIR and, to a lesser extent, the SIR showed a pattern similar to the C1 ORF region (data not shown). The C1 intron, however, had an extreme over-representation of the 24 nt size class with the other size classes being either nearly (22 nt) or totally (21 and 23 nt) absent (Figure 2).

Since the VARX plant was also infected with SSEV, a similar analysis of SSEV-derived siRNAs was performed. Mapping against the genome of SSEV (NC_001868) demonstrated that 0.17% of total reads (6572) were derived from it and that these reads covered 98.6% of the SSEV genome at an average depth of 55X, leaving only 4 gaps of between 5 and 15 nucleotides (Figure S1). Although showing some high degrees of local heterogeneity, genome coverage of SSEV was less biased when comparing the different genomic regions. Nevertheless a similar trend to that associated with SWSV was observed with a higher depth of coverage for the virion sense V1–V2 ORFs (76.5X) than for both the complementary sense C1 ORF (37.6X) and the non-coding regions (46X). Also, as for SWSV, the 21–22 nt siRNA size classes were enriched amongst those mapping to the virion sense ORFs and the 24 nt, siRNA size class was enriched amongst those mapping to the complementary sense C1 ORF (Figure S1). However, unlike for SWSV, no strong siRNA size-class biases were observed for the non-coding regions (data not shown).

By collectively using the Illumina siRNA reads and the 454 VANA reads it was possible to assemble a single genome of the novel mastrevirus from both the VARX and USDA plants.

SWSV

Associations between siRNAs and SWSV/SSEV genomic and transcript secondary structures

It has been previously determined that nucleic acid structures can have an appreciable impact on both the distribution of siRNA targets [36,37], and the operational efficiency of small RNA mediated anti-viral and anti-viroid defences [37,38]. We detected strong evidence for the presence of ssDNA secondary structures in both the SWSV (30 high confidence structure set (HCSS) identified) and SSEV (29 HCSS structures identified) genomes (Table 3). The distributions of the HCSS structural elements were, however, different in the predicted virion and complementary strand transcripts of the two viruses, with only two HCCS structures detected in the SWSV complementary strand transcript and none being detected in the SSEV virion strand transcript (so that this particular transcript was not analysed further).

We detected a strong association between the absence of predicted secondary structures within the ssDNA SWSV genome and increased frequencies of corresponding 22, 23 and 24 nt long siRNAs (p-values <0.008; Table 3). Curiously, we found a

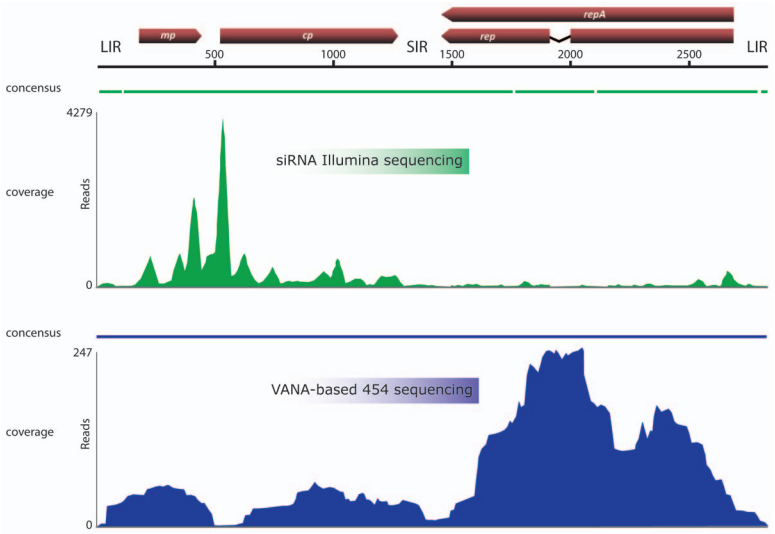


Figure 1. SWSV genome coverage following NGS. The genomic organization of SWSV is schematically shown above the graph. While relative degrees of coverage achieved after *a posteriori* mapping of reads produced by Illumina-based siRNA sequencing against the SWSV genome is indicated in green, the coverage achieved after mapping reads produced by 454 GS FLX Titanium-based VANA sequencing is indicated in blue.
doi:10.1371/journal.pone.0102945.g001

different association when considering the predicted SWSV RNA transcripts with 21 nt siRNA reads displaying a strong tendency to correspond with nucleotide sites that were predicted to be base paired in both the virion and complementary strand transcripts (p-values $<2.49\times10\text{-}6$) and the 22, 23 and 24 siRNA size classes displaying a similar tendency with respect to the virion strand transcript (p-values $<6.07\times10\text{-}13$).

Similar to SWSV, for the SSEV full genome there was an association between the absence of ssDNA structural elements and increased frequencies of 22 nt siRNAs. Also similar to SWSV the 22, 23 and 24 nt long siRNAs display a significant tendency to correspond with transcript nucleotides that are base-paired within secondary structural elements.

A novel sugarcane-infecting mastrevirus originating from the Nile region

The complete genome of SWSV, as recovered from the VARX and USDA plants, is most similar to that of *Wheat dwarf India virus* (WDIV, Accession number NC_017828), with which it shares 61% genome-wide identity. Whereas the Rep and MP amino acid sequences of SWSV are also most similar to those of WDIV (54.4% and 44.8% identity, respectively), the CP is most similar to that of *Panicum streak virus* (PanSV, NC_001647, 51.4–53.9%). Based on the 78% species demarcation threshold set by the Geminivirus study group of the ICTV [32], it is clear that the novel mastrevirus should be considered a new species within the genus *Mastrevirus* of the Family *Geminiviridae* (Figure S2). This is further confirmed by phylogenetic analyses performed on both the full genome (Figure 3) and on the amino acid sequences of its encoded proteins (Figure 4). The new virus clearly clusters with mastreviruses on a branch that is not closely associated with any other species classified within this genus. Whereas the CP of SWSV clusters within the virus clade including the various African streak viruses, Australasian striate mosaic viruses, *Digitaria streak virus* (DSV) and WDIV, the Reps cluster with the African streak viruses and WDIV (Figure 4).

SWSV was not detected in sugarcane seedlings derived from sugarcane true seeds under sterile insect-proof conditions, in agreement with the fact that seed transmission of geminiviruses has not so far been reported. The novel mastrevirus was, however, detected in five sugarcane plants originating from Sudan (A0037, B0065, B0069, D0005 and E0144) out of the 23 screened (Table S1).

Complete SWSV genomes from four sugarcane plants (A0037, B0069, D0005 and E0144) were cloned and sequenced. The genomes of these isolates have >91% genome-wide identity with those recovered from the VARX and USDA sugarcane plants. Phylogenetic analyses of the full genomes (Figure 3) and of the amino acid sequences that they likely encode (Figure 4) confirmed that the isolates obtained from the Sudanese sugarcane plants also belong to the SWSV species. The six isolates can be further classified into 3 strains, SWSV-A (VARX, USDA), -B (B0069, D0005, E0144) and -C (A0037) (Figure S3) based on the proposed classification of mastrevirus strains outlined by Muhire et al. [32].

Using primers that allow the amplification of all sugarcane-infecting mastreviruses, including Sugarcane streak Egypt Virus, Sugarcane streak virus, Maize streak virus, Sugarcane streak Reunion virus, Eragrostis streak virus and Saccharum streak virus, the five sugarcane plants originating from Sudan were shown to be free of co-infection with other known mastreviruses. Three of them are still maintained at the CIRAD sugarcane quarantine station (B0065, B0069 and D0005) and exhibit white spots on the base of their leaf blades, around the blade joint where the two wedge shaped areas called "dewlaps" are located (Figure S4). These spots can become fused laterally, so as to appear as chlorotic stripes (Figure S4). It is noteworthy that the SWSV infected D0005 plant displayed very little evidence of these spots (only one leaf out of eight exhibited tiny white spots that resembled thrip damage) and it is therefore very likely that SWSV infections could escape visual inspection (Figure S4). Given that three of the infected sugarcane varieties exhibited mild foliar symptoms, i.e. white spots on the base of their leaf blades that become fused laterally, so as to appear as chlorotic stripes, we propose naming the new species Sugarcane white streak Virus.

Genome analysis of SWSV

The SWSV genomes recovered from the various sugarcane plants were between 2828 and 2836 nt and are, in almost all

Table 2. Lengths, numbers of reads and BlastX analysis results for siRNA *de novo* contigs from sugarcane plant VARX with detectable homology to mastreviral sequences.

Virus	Contig	Contig length (bp)	Number of reads	BlastX Virus	BlastXLocus	BlastX e-value	Percent identity
SSEV	#121	101	270	SSEV (AAF76871)	CP	2.60E–9	100%
	#176	133	520	SSEV (AAC98080)	MP	7.22E–15	100%
SWSV	#44	117	914	SacSV (YP_003288767)	CP	7,16E–06	68%
	#79	275	7640	WDIV (YP_006273068)	MP	3,29E–11	70%
	#86	258	1649	WDIV (YP_006273069)	CP	6,78E–20	50%
	#101	111	1284	SSRV (ABZ03975)	CP	1,94E–05	64%

Acronyms used are as follows: SSEV (Sugarcane streak Egypt virus), SWSV (Sugarcane white streak virus), SacSV (Saccharum streak virus), WDIV (Wheat dwarf India virus), SSRV (Sugarcane streak Reunion virus).

doi:10.1371/journal.pone.0102945.t002

respects, very similar to those of all other previously described mastreviruses. The one exceptional feature of the SWSV genomes is that in case of the VARX and USDA isolates alternative splicing of complementary sense transcripts likely results in the expression of both a standard Rep (which is predicted to be 396 amino acids long), and a rather unusual RepA of 418 amino acids long. This is the only known occurrence in any geminivirus of a RepA that is larger than Rep.

Given the uniqueness of this apparent genome organisation in the USDA and VARX isolates the correct identification of the intron within the complementary sense transcript was verified. RT-PCR reactions targeting the complementary sense transcript clearly indicated the presence of a mixture of spliced and non-spliced complementary sense mRNA transcripts, and confirmed that the correct locations of the acceptor and donor sites of the 66 nt long SWSV intron had been identified (Figure S5).

Analysis of recombination

All the SWSV genome sequences determined here share evidence of the same ancestral recombination event in the short intergenic region - corresponding to genomic coordinates 1419–1468 in the USDA isolate ($p = 3.821 \times 10\text{-}7$ for the GENECONV, MAXCHI and RDP methods implemented in RDP4.24). Corresponding coordinates are known to be very common sites of recombination in mastreviruses [39] and the fragment that they delimit in SWSV has apparently been derived from something resembling an African streak virus.

Discussion

We have performed NGS-based analyses of both siRNA and VANA isolated from sugarcane plants originating from Egypt. Both sequence-independent NGS approaches revealed the presence of a novel mastrevirus, SWSV, which had so far escaped routine quarantine detection assays, possibly because it was present in mixed infection with SSEV. The procedures used for the discovery of SWSV pave the way towards the application of NGS-based quarantine detection procedures. Such procedures would likely be hierarchical with a first stage sequence-independent NGS step followed by sequence-dependent secondary assays. Whereas the first step would be to identify novel viruses within a single plant (perhaps one displaying apparent disease symptoms), the second step would be to use sequence dependent approaches to both confirm the presence of any novel virus(es) identified in the original host, and identify the presence of this(ese) virus(es) in larger plant collections. A major strength of such an approach is that it would also yield complete genome sequences.

The present study also confirms that both VANA [13] and siRNA [16] can be successfully targeted by metagenomics approaches for the discovery and characterization of plant-infecting DNA viruses. The VANA-based 454 pyrosequencing approach has several advantages as it initially combines reverse-transcriptase priming and a Klenow Fragment step, which potentially enables the detection of both RNA and DNA viruses. Additionally, up to 96-tagged amplified DNAs (cDNA and DNAs amplified using the Klenow Fragment step) can be pooled and sequenced in multiplex format [15] making this approach very useful for routine diagnostic screening of plants within BRCs and quarantine stations. However, validation using plants infected or co-infected with RNA and DNA viruses needs to be carried out in order to determine the sensitivity and specificity levels of this 454-based VANA sequencing approach.

Virus-derived siRNAs naturally accumulate in virus-infected plants as a consequence of the action of Dicer enzymes as part of

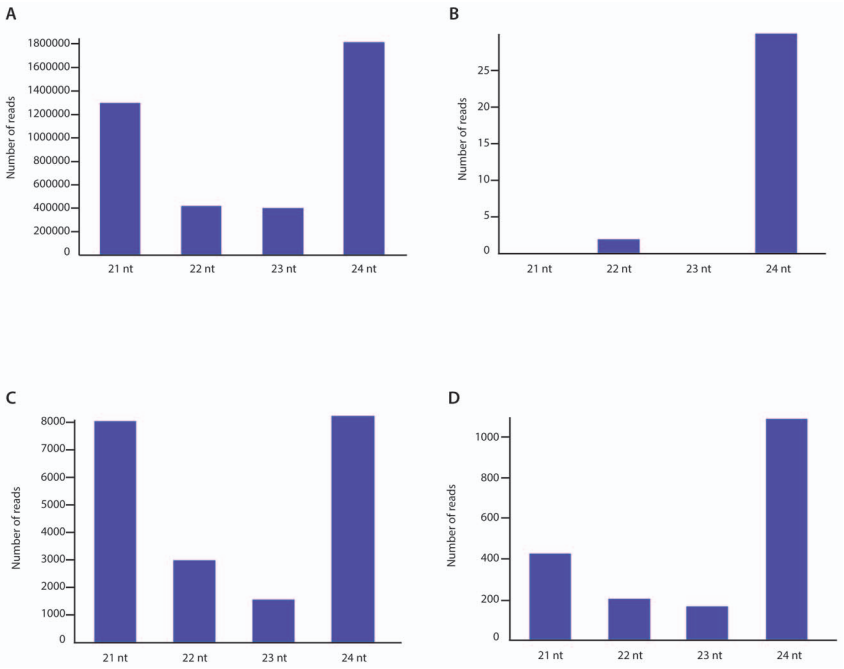


Figure 2. Size distribution of sequenced siRNAs obtained from the VARX plant. The histograms represent the numbers of siRNA reads in each size class. (A) The size distributions of total reads, (B) The size distributions of reads mapping to the *rep* gene C-sense intronic region of SWSV, (C) The size distributions of reads mapping to the V1–V2 ORFs region of SWSV and (D) The size distributions of reads mapping to the C1 ORF region of SWSV.
doi:10.1371/journal.pone.0102945.g002

the RNA silencing-based plant antiviral defences [40]. Adopting a metagenomic approach and randomly sequencing these siRNAs is therefore an extremely powerful way to discover and characterise previously unknown plant viruses and viroids [16,41]. In addition to providing evidence for the presence of the two mastreviruses co-infecting the VARX sugarcane plant, this approach provided information on the interaction of the plant antiviral silencing machinery and these two viruses. Although these aspects have been studied previously in geminiviruses in the *Begomovirus* genus (Blevins et al., 2006; Akbergenov et al., 2006; Rodríguez-Negrete et al., 2009; Yang et al., 2011; Aregger et al., 2012), very little comparable information has previously been available for mastreviruses.

The siRNA distributions observed here for SWSV and SSEV, perhaps unsurprisingly, seem to largely parallel those previously reported for begomoviruses. In particular, the differences in size classes observed between different genome regions suggest that mastreviruses are subject both to transcriptional gene silencing, based on 24 nt long siRNAs produced through the action of DCL3, and to post-transcriptional gene silencing (PTGS) mediated by the 21–22 nt long siRNAs produced through the action of the antiviral Dicers DCL4 and DCL2 (Rodríguez-Negrete et al., 2009; Aregger et al., 2012). The action of the former mechanism is particularly evident in the siRNAs mapping to the SWSV intron but is also, to a lesser extent, evident in the siRNAs mapping to both the non-coding regions and the complementary sense ORFs of SWSV and SSEV. On the other hand, the 21–22 nt siRNA size classes associated with PTGS are particularly evident in the siRNAs mapping to the two virion sense ORFs which are known to be more actively transcribed in mastreviruses than their complementary sense counterparts [42].

For both SWSV and SSEV we detected a significant association between the frequencies of siRNAs and the presence/absence of predicted secondary structures within both the single stranded DNA (ssDNA) genomes of these viruses and their predicted single stranded RNA (ssRNA) complementary and virion strand transcripts. However, whereas significantly more siRNAs corresponded with unstructured regions of the ssDNA genome, for the transcripts significantly more siRNAs corresponded with structured regions of ssRNA. It is plausible that base-paired nucleotides within transcript RNA molecules are protected from siRNA binding and that the secondary structures evident both in transcripts produced by SWSV, SSEV and in mastrevirus genomes in general [43] may represent an evolutionary adaptation for viral persistence. In mammalian RNA viruses there is an association between degrees of genomic secondary structure and infection duration with viruses having highly structured genomes tending to cause chronic infections and viruses with unstructured genomes tending to cause acute infections [44,45].

In both analysed Egyptian sugarcane accessions, VARX and USDA, SWSV was found to be present in co-infections with SSEV. Both sugarcane plants were independently collected in Egypt which suggests that SWSV infection of Egyptian sugarcane plants may not be a rare phenomenon. SWSV was also detected in SSEV-free plants that originated from Sudan. It is noteworthy that one of the Sudanese plants from which SWSV was isolated, E0144, was initially grown in Sudan in 1992 before being transferred to Barbados in 1998 and subsequently sent back to the CIRAD Sugarcane Quarantine Station in 2009 (unpublished data, CIRAD Sugarcane Quarantine Station). Assuming that SWSV did not infect this plant in Barbados between 1998 and 2009, it is plausible that SWSV was present along the Nile basin at least from the late 1980s. Interestingly, as a consequence of indel polymorphisms in the 66 nt long SWSV intron, the Egyptian SWSV isolates VARX and USDA have a highly unusual genome organization and likely express a RepA protein that, while having

Table 3. Associations between siRNAs and SWSV/SSEV genomic and transcript secondary structures in the HCSS.

Sequence name	Component	Length	Number of structures	siRNA type	Probability if association between siRNAs and secondary structure (KS Test)	Probability of no association between siRNAs and base-paired nucleotides (WRS test)	Probability of no association between siRNAs and unpaired nucleotides (WRS test)
SWSV	Full genome	2830	30	All	**5.69×10^{-5}**	0.999	**6.20×10^{-4}**
				21	0.205	0.590	0.410
				22	**3.35×10^{-6}**	0.992	**0.008**
				23	**0.0005**	0.999	**6.81×10^{-4}**
				24	**3.48×10^{-9}**	0.999	**2.67×10^{-6}**
	V-strand transcript	1222	16	All	**0**	**1.80×10^{-15}**	**1**
				21	**0.022**	**4.93×10^{-17}**	**1**
				22	**0**	**4.1×10^{-17}**	**1**
				23	**8.23×10^{-14}**	**1.10×10^{-16}**	**1**
				24	**0**	**6.07×10^{-13}**	**1**
	C-strand transcript	1446	2	All	**2.00×10^{-4}**	0.100	0.899
				21	**1.78×10^{-7}**	**2.49×10^{-6}**	0.999
				22	0.409	0.122	0.877
				23	**0.008**	**0.925**	0.075
				24	**0.003**	**0.970**	**0.029**
SSEV	Full genome	2706	29	All	**0.007**	**0.889**	0.111
				21	**0.016**	**0.737**	0.263
				22	**0.0001**	**0.999**	**0.001**
				23	0.140	0.861	0.139
				24	0.209	0.499	0.501
	V-strand transcript	1131	0	NA	NA	NA	NA
	C-strand transcript	1406	13	All	**0.002**	**0.019**	0.981
				21	0.670	0.465	0.535
				22	**0.006**	**0.025**	0.975
				23	**0.002**	**0.001**	0.999
				24	**0.041**	**0.023**	0.977

doi:10.1371/journal.pone.0102945.t003

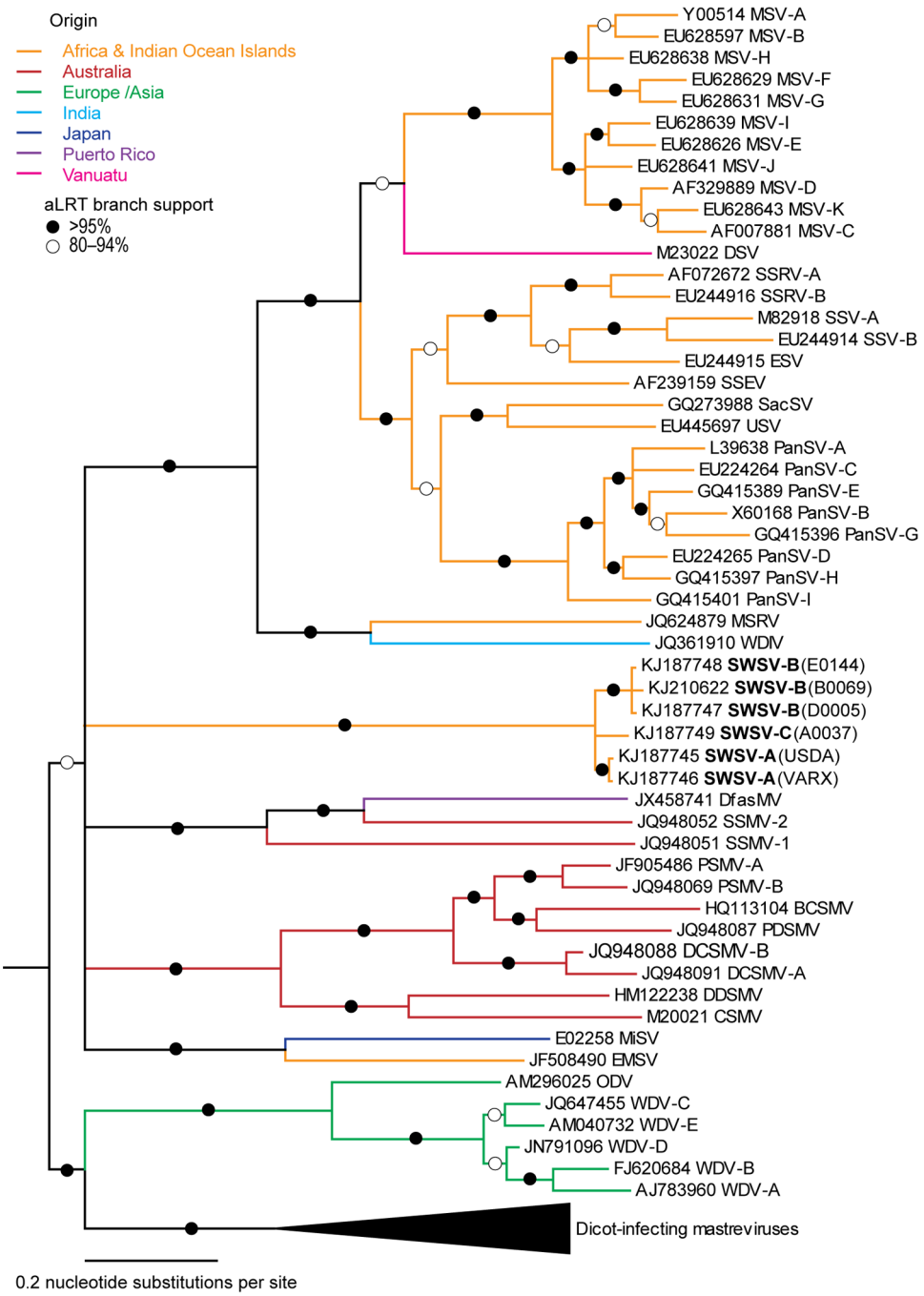


Figure 3. Maximum-likelihood phylogenetic tree of 63 virus isolates representing each known mastrevirus species (including major strains) and the 6 SWSV isolates determined in this study. Tree branches are coloured according to the geographical origins of the viruses. Branches marked with filled and open circles respectively have >95% and 80–94% approximate likelihood ratio test support; branches having <80% support were collapsed. The phylogenetic tree is rooted using the full genome sequence of Dicot-infecting mastreviruses.
doi:10.1371/journal.pone.0102945.g003

the same N- and C-terminus sequences as Rep, is 22 amino acids longer than Rep.

The recent discoveries of SWSV and other highly divergent mastreviruses [46,47] suggest that this geminivirus genus likely encompasses a far greater diversity and has a greater global distribution than has been previously appreciated. The SWSV isolate from the Sudanese sugarcane plant that had been propagated in Barbados represents only the third instance of discovery of mastreviruses in the New World [16,48], and suggests that there may have been other undetected recent introductions of mastreviruses to the Americas. Although insect transmission of mastreviruses in the New World remains to be reported, it is noteworthy that one of the three mastrevirus species that has so far been detected in the Americas was isolated from a dragonfly which had possibly eaten a plant feeding insect that was carrying the virus [48]. The presence of SWSV in Barbados offers an opportunity to investigate possible natural transmission of the virus by screening sugarcane planted near the SWSV infected

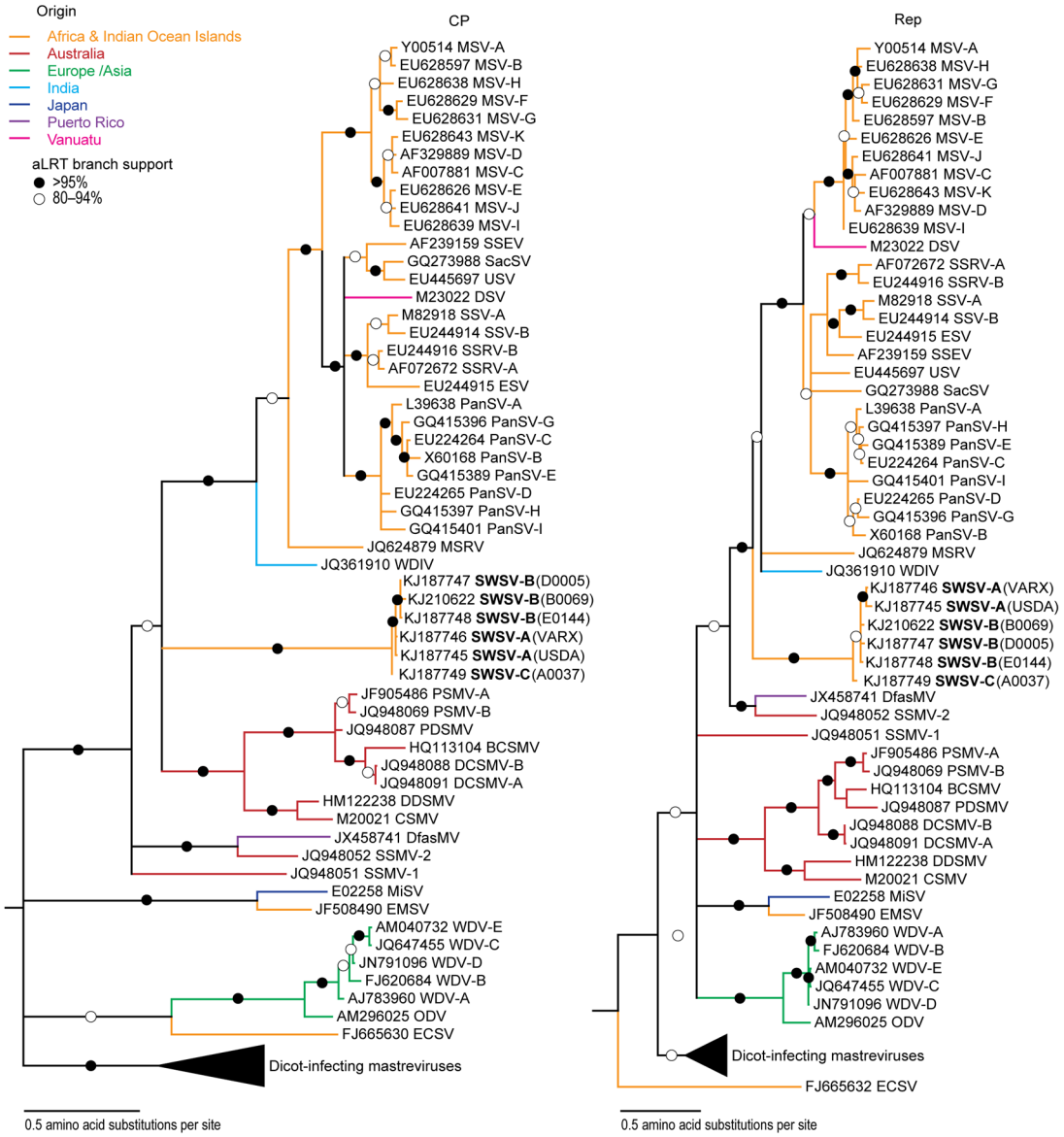


Figure 4. Maximum-likelihood phylogenetic tree of Rep (A) and CP (B) proteins. Tree branches are coloured according to the geographical origins of the viruses. Branches marked with filled and open circles are respectively have >95% and 80–94% approximate likelihood ratio test support; branches having <80% support were collapsed.
doi:10.1371/journal.pone.0102945.g004

E0144 accessions. Phylogenetic analyses of any SWSV genomes sampled from such plants should reveal their likely recent transmission histories.

Given the relatively high degrees of sequence divergence observed between the different SWSV isolates described here (~9%), it is plausible that the natural geographical range of SWSV is broader than just the Nile basin. Also, the global dissemination of sugarcane cuttings, the absence of SWSV diagnostic tools, and the fact that SWSV induces, at least in one case, extremely mild symptoms in sugarcane imply that SWSV may have already been unknowingly distributed throughout the sugarcane growing regions of the world. The failure of established sugarcane quarantine diagnostics in this regard provides a dramatic example of how potentially pathogenic viruses can evade the screening procedures of quarantine facilities and may spread worldwide through international plant material exchanges. In this regard the situation with SWSV might closely match that of *Sugarcane yellow leaf virus* (SCYLV), which remained unnoticed for at least 30 years during its spread throughout the world [49]. In order to accurately determine the potential economic impacts of the dissemination of SWSV, additional studies assessing the pathogenicity of this virus are certainly warranted.

Our study stresses both the potential advantages of NGS-based virus metagenomic screening in a plant quarantine setting, and the need to better assess viral diversity within plants that are destined for exotic habitats. It indicates that a combination of sequence-independent NGS-based partial viral genome sequencing coupled with sequence-dependent Sanger-based full genome cloning and sequencing is likely to reduce the number of non-intercepted virus pathogens passing through plant quarantine stations, while at the same time alerting authorities to the presence and potential spread of viruses with unknown pathogenic potentials.

Supporting Information

Figure S1 (A) Genome coverages obtained after *a posteriori* mapping against the complete genome of SSEV of reads produced by Illumina (siRNA sequencing). The genomic organization of SSEV is schematically shown at the top of the figure. (B) Size distribution of sequenced siRNAs obtained from the VARX plant mapping on the V1–V2 ORFs region of SSEV. Histograms represent the number of siRNA reads in each size class. and (C) Size distribution reads mapping on C1–C2 ORFs region of SSEV.
(TIF)

Figure S2 Two-dimensional genome-wide percentage pairwise nucleotide identity plot of monocot-infecting mastreviruses including the six novel SWSV isolates from this study.
(TIF)

Figure S3 (A) Maximum-likelihood phylogenetic tree of six SWSV isolates. The six isolates can be classified into 3 strains, SWSV-A (VARX, USDA), -B (B0069, D0005, E0144) and -C (A0037). (B) Genome-wide pairwise nucleotide similarity score matrix, the 94% strain demarcation threshold set by the Geminivirus study group of the ICTV (Muhire et al. 2013) is indicated (green coloured below 94% and pink-red coloured above 94%).
(TIF)

Figure S4 Symptoms caused by SWSV on B0065, B0069 and D0005 plants.
(TIF)

Figure S5 Reverse transcriptase priming and amplification of nucleic acids were carried out in order to detect the *rep* gene C-sense intronic region. (A) Agarose gel detection of presence of a mixture of spliced and non-spliced complementary sense mRNA transcripts. 1: 1 Kb ladder; 2: Reverse transcriptase priming and amplification of nucleic acids without DNase treatment of extracted RNAs; 3: Reverse transcriptase priming and amplification of nucleic acids with DNase treatment of extracted RNAs. (B) 66 nt long SWSV intron nucleotidic sequence and splice donor and acceptor sites. The sequence of the intron (in lower case) and its flanking exons (upper case) are shown. The 5′ (donor) and 3′ (acceptor) splice sites are underlined (lower case).
(TIF)

Table S1 List of the sugarcane varieties from the CIRAD Sugarcane Quarantine Station (SQS) that were screened for the presence of all known sugarcane-infecting mastreviruses and SWSV.
(DOC)

Acknowledgments

The authors are grateful to Sébastien Theil for depositing NGS datasets in GenBank.

Author Contributions

Conceived and designed the experiments: TC DF PR. Performed the experiments: BM CJ SG GF JHD EF PB. Analyzed the data: TC DF BM DPM AV PR. Wrote the paper: TC DPM AV PR.

References

1. Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, et al. (2004) Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol 19: 535–544.
2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. (2008) Global trends in emerging infectious diseases. Nature 451: 990–993.
3. Jones RAC (2009) Plant virus emergence and evolution: Origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Research 141: 113–130.
4. Roossinck MJ (2011) The good viruses: viral mutualistic symbioses. Nat Rev Microbiol 9: 99–108.
5. Rosario K, Breitbart M (2011) Exploring the viral world through metagenomics. Curr Opin Virol 1: 1–9.
6. van der Heijden MG, Bardgett RD, van Straalen NM (2008) The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett 11: 296–310.
7. Li L, Delwart E (2011) From orphan virus to pathogen: the path to the clinical lab. Curr Opin Virol 1: 282–288.
8. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2: 63–67.
9. Willner D, Hugenholtz P (2013) From deep sequencing to viral tagging: recent advances in viral metagenomics. Bioessays 35: 436–442.
10. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. NatProtoc 4: 470–483.
11. Lipkin WI (2010) Microbe hunting. Microbiol Mol Biol Rev 74: 363–377.
12. Roy A, Shao J, Hartung JS, Schneider W, Brlansky RH (2013) A case study on discovery of novel *Citrus Leprosis* virus cytoplasmic type 2 utilizing small RNA libraries by Next Generation Sequencing andbioinformatic analyses. Journal of Data Mining in Genomics & Proteomics 4: 1000129.
13. Melcher U, Muthukumar V, Wiley GB, Min BE, Palmer MW, et al. (2008) Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. Journal of Virological Methods 152: 49–55.
14. Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL (2008) Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog 4: e1000163.
15. Roossinck MJ, Saha P, Wiley G, Quan J, White J, et al. (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19 81–88.
16. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, et al. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388: 1–7.
17. Al Rwahnih M, Daubert S, Golino D, Rowhani A (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387: 395–401.
18. Kraberger S, Stainton D, Dayaram A, Zawar-Reza P, Gomez C, et al. (2013) Discovery of Sclerotinia sclerotiorum Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. Genome Announc 1.
19. Li R, Gao S, Hernandez AG, Wechter WP, Fei Z, et al. (2012) Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLoS One 7: e37127.
20. Martinez G, Donaire L, Llave C, Pallas V, Gomez G (2010) High-throughput sequencing of Hop stunt viroid-derived small RNAs from cucumber leaves and phloem. Mol Plant Pathol 11: 347–359.
21. Sikorski A, Massaro M, Kraberger S, Young LM, Smalley D, et al. (2013) Novel myco-like DNA viruses discovered in the faecal matter of various animals. Virus Res 177: 209–216.
22. Borucki MK, Chen-Harris H, Lao V, Vanier G, Wadford DA, et al. (2013) Ultra-Deep Sequencing of Intra-host Rabies Virus Populations during Cross-species Transmission. PLoS Negl Trop Dis 7: e2555.
23. Simmons HE, Dunham JP, Stack JC, Dickins BJ, Pagan I, et al. (2012) Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J Gen Virol 93: 1831–1840.
24. Wang H, Xie J, Shreeve TG, Ma J, Pallett DW, et al. (2013) Sequence recombination and conservation of Varroa destructor virus-1 and deformed wing virus in field collected honey bees (Apis mellifera). PLoS One 8: e74508.
25. Bigarre L, Salah M, Granier M, Frutos R, Thouvenel J, et al. (1999) Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. Arch Virol 144: 2331–2344.
26. Shamloul AM, Abdallah NA, Madkour MA, Hadidi A (2001) Sensitive detection of the Egyptian species of sugarcane streak virus by PCR-probe capture hybridization (PCR-ELISA) and its complete nucleotide sequence. J Virol Methods 92: 45–54.
27. Froussard P (1993) rPCR: a powerful tool for random amplification of whole RNA sequences. PCR Methods Appl 2: 185–190.
28. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
29. Posada D (2008) jModelTest: Phylogenetic model averaging. Molecular Biology and Evolution 25: 1253–1256.
30. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.

31. Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 537: 113–137.
32. Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, et al. (2013) A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Archives of Virology 158: 1411–1424.
33. Martin DP, Lemey P, Lott M, Moulton V, Posada D, et al. (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26: 2462–2463.
34. Semegni JY, Wamalwa M, Gaujoux R, Harkins GW, Gray A, et al. (2011) NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27: 2443–2445.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403–410.
36. Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, et al. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392: 203–214.
37. Itaya A, Zhong X, Bundschuh R, Qi Y, Wang Y, et al. (2007) A structured viroid RNA serves as a substrate for dicer-like cleavage to produce biologically active small RNAs but is resistant to RNA-induced silencing complex-mediated degradation. J Virol 81: 2980–2994.
38. Westerhout EM, Ooms M, Vink M, Das AT, Berkhout B (2005) HIV-1 can escape from RNA interference by evolving an alternative structure in its RNA genome. Nucleic Acids Res 33: 796–804.
39. Varsani A, Monjane AL, Donaldson L, Oluwafemi S, Zinga I, et al. (2009) Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. Virol J 6: 194.
40. Voinnet O (2005) Induction and suppression of RNA silencing: insights from viral infections. Nat Rev Genet 6: 206–220.
41. Kashif M, Pietila S, Artola K, Jones RAC, Tugume AK, et al. (2012) Detection of Viruses in Sweetpotato from Honduras and Guatemala Augmented by Deep-Sequencing of Small-RNAs. Plant Disease 96: 1430–1437.
42. Morris-Krsinich BA, Mullineaux PM, Donson J, Boulton MI, Markham PG, et al. (1985) Bidirectional transcription of maize streak virus DNA and identification of the coat protein gene. Nucleic Acids Res 13: 7237–7256.
43. Muhire BM, Golden M, Murrell B, Lefeuvre P, Lett JM, et al. (2013) Evidence of pervasive biologically functional secondary-structures within the genomes of eukaryotic single-stranded DNA viruses. J Virol.
44. Davis M, Sagan SM, Pezacki JP, Evans DJ, Simmonds P (2008) Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. J Virol 82: 11824–11836.
45. Simmonds P, Tuplin A, Evans DJ (2004) Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA 10: 1337–1351.
46. Kraberger S, Harkins GW, Kumari SG, Thomas JE, Schwinghamer MW, et al. (2013) Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. Virology 444: 282–291.
47. Kraberger S, Thomas JE, Geering AD, Dayaram A, Stainton D, et al. (2012) Australian monocot-infecting mastrevirus diversity rivals that in Africa. Virus Res 169: 127–136.
48. Rosario K, Padilla-Rodriguez M, Kraberger S, Stainton D, Martin DP, et al. (2013) Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. Virus Res 171: 231–237.
49. Komor E, ElSayed A, Lehrer AT (2010) Sugarcane yellow leaf virus introduction and spread in Hawaiian sugarcane industry: Retrospective epidemiological study of an unnoticed, mostly asymptomatic plant disease. European Journal of Plant Pathology 127: 207–217.

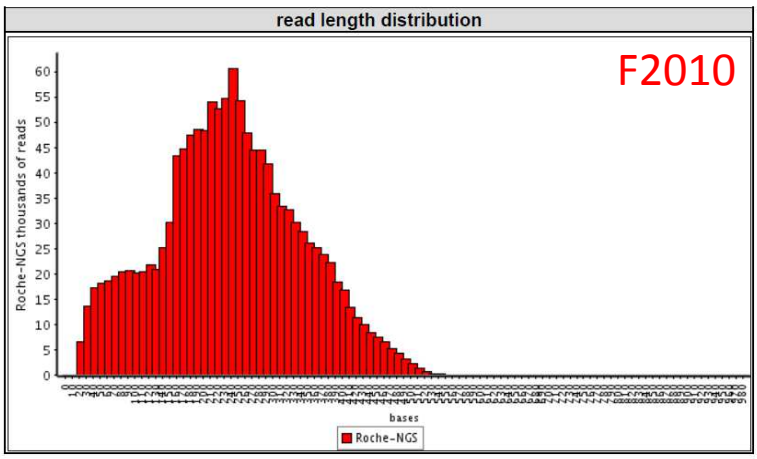
Annexe 3 :

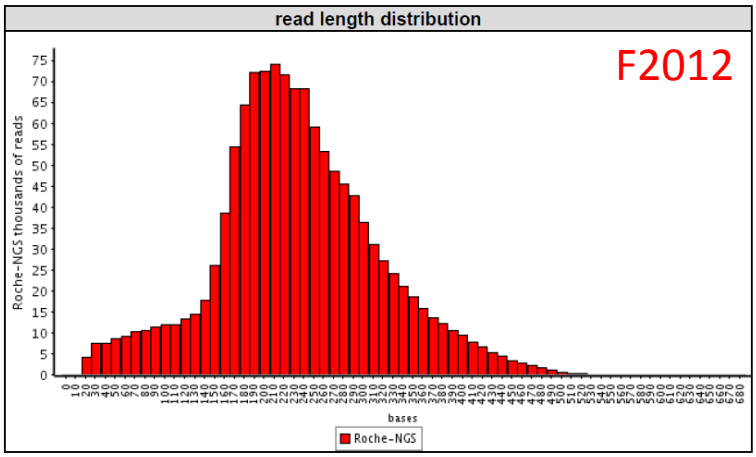
Informations générales sur les reads issus du pyroséquençage pour chaque période d'échantillonnage.

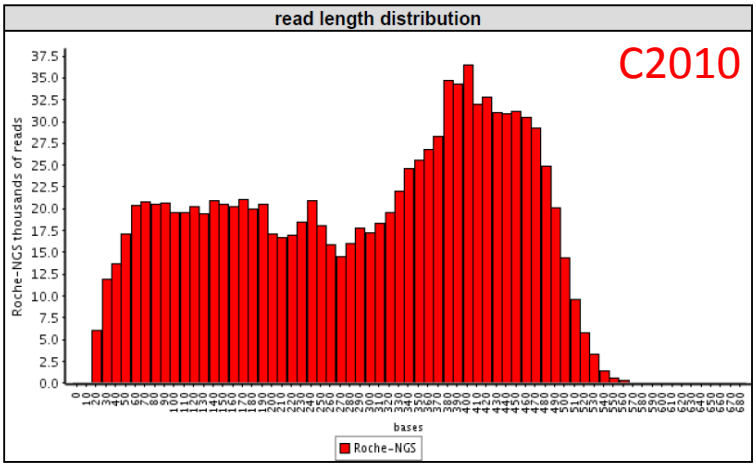
	Huitième de plaque	Nombre de bases (Mb)	Nombre de reads	Longueur moyenne des reads (bp)	Longueur médiane des reads (bp)
F20100	1	36,5	152482	239	234
	2	32,9	129133	255	254
	3	36,9	151056	244	243
	4	45,4	185301	245	244
	5	43,7	181707	240	237
	6	48,4	203218	238	237
	7	38,8	146692	264	267
	8	45,4	183035	248	249
F20120	1	34,4	146696	234	228
	2	44,7	183167	244	240
	3	38,1	152341	250	246
	4	41,3	167242	247	243
	5	26,9	110326	244	235
	6	39,3	158845	247	241
	7	37,8	160012	236	229
	8	43,1	181138	238	232
C20100	1	33,4	123423	271	258
	2	36,8	120078	306	332
	3	37,5	127049	295	318
	4	49,1	167579	293	312
	5	49,5	160822	308	338
	6	50	158348	316	346
	7	30,6	101533	301	334
	8	41,9	133519	314	355
C20120	1	45,9	176413	260	256
	2	42,5	156380	272	267
	3	61,6	264165	233	230
	4	44,9	164816	272	272
	5	39	147362	265	260
	6	47,7	181189	263	260
	7	44,1	162152	272	266
	8	50,6	192474	263	257

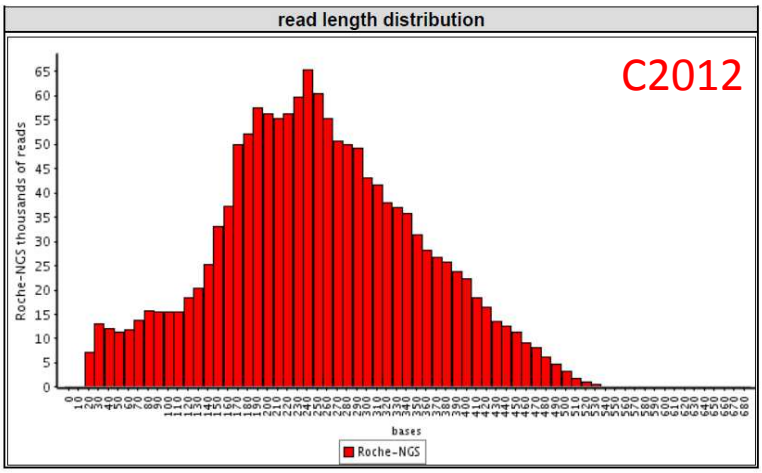
Annexe 4 :

Distribution de la taille des reads issus du pyroséquençage pour chaque période d'échantillonnage.


read length distribution
F2010
Roche-NGS thousands of reads
bases
Roche-NGS


read length distribution
F2012
Roche-NGS thousands of reads
bases
Roche-NGS


read length distribution
C2010
Roche-NGS thousands of reads
bases
Roche-NGS


read length distribution
C2012
Roche-NGS thousands of reads
bases
Roche-NGS

Annexe 5 :

Tableau répertoriant les nouvelles espèces virales potentielles découvertes via nos travaux de géo-métagénomique.

F2010				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
1A-014	Euphorbia caput-medusae	Geminiviridae	2	66,67-68,33
1A-014	Euphorbia caput-medusae	Secoviridae	2	37,5
1A-014	Euphorbia caput-medusae	Endornaviridae	1	45,83
1A-019	Nylandtia spinosa	Luteoviridae	1	35,62
1A-038	Cynodon dactylon	Geminiviridae	2	24,10-32,10
1A-047	Briza maxima	Geminiviridae	1	36,05
1A-049	Briza maxima	Partitiviridae	20	43,33-50
1A-070	Inconnu	Geminiviridae	36	33,08-52,63
1A-077	Avena byzantina	Nanoviridae	2	42,86-43,3
1A-077	Avena byzantina	Geminiviridae	1	44,3
1A-086	Avena byzantina	Partitiviridae	1	49,04
1A-092	Ixia dubia	Partitiviridae	4	43,24-60
1A-099	Bromus diandrus	Partitiviridae	2	50,53-43,28
1A-099	Bromus diandrus	Geminiviridae	2	35,29-43,28
1B-004	Conicosia pugioniformis	Endornaviridae	2	35,29
1B-007	Avena byzantina	Bromoviridae	2	45,1
1B-007	Avena byzantina	Betaflexiviridae	1	37,18
1B-007	Avena byzantina	Tymoviridae	1	32,86
1B-012	Galenia africana	Partitiviridae	1	45,24
1B-051	Capnophyllum africana	Partitiviridae	10	35,04-60,90
1B-051	Capnophyllum africana	Tymoviridae	1	45,16
1B-068	Inconnu	Partitiviridae	10	55,05-66,07
1B-079	Briza maxima	Nanoviridae	8	38,46-53,97
1B-085	Rafnia angulata	Caulimoviridae	1	54,76
1B-089	Cysticapnos vesicaria	Secoviridae	4	40,54-47-27
1B-089	Cysticapnos vesicaria	Geminiviridae	1	35,29
1B-096	Avena byzantina	Partitiviridae	75	34,29-62,16
1B-098	Avena byzantina	Geminiviridae	1	35,29
1C-017	Gazania pectinata	Virgaviridae	3	36,19-45,10
1C-024	Indigofera heterophylla	Alphaflexiviridae	1	58,33
1C-052	Bromus pectinatus	Geminiviridae	1	44,83
1C-059	Lolium rigidum	Caulimoviridae	1	33,33
1C-091	Stachys aethiopica	Partitiviridae	2	42,17
1C-093	Phalaris minor	Caulimoviridae	1	38,81
1C-093	Phalaris minor	Tymoviridae	4	35,14-40,95
1C-094	Phalaris minor	Partitiviridae	2	38,3-43,66
1D-006	Rhus laevigata	Umbravirus	1	44,09
1D-013	Staberoha distachyos	Partitiviridae	3	59,78
1D-028	Inconnu	Totiviridae	33	64,81-31,51
1D-028	Inconnu	Geminiviridae	19	29,89-63,33
1D-041	Willdenowia incurvata	Tymoviridae	2	34,43-38,20
1D-053	Cyclopia genistoides	Geminiviridae	4	37,29-47,17
1D-057	Clutia alaternoides	Tombusviridae	137	50-96,65
1D-068	Ehrharta calycina	Partitiviridae	1	42,86

F2010 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
1D-068	Ehrharta calycina	Tombusviridae	5	48,86-54,05
1D-068	Ehrharta calycina	Umbravirus	3	46,24-53,62
1D-087	Bromus diandrus	Partitiviridae	3	37,93-56,90
1D-098	Avena byzantina	Geminiviridae	4	46,48-62,16
1E-028	Ammocharis longifolia	Geminiviridae	2	32,48-57,58
1E-028	Ammocharis longifolia	Partitiviridae	3	22,43-60,47
1E-050	Briza maxima	Partitiviridae	2	55,56
1E-050	Briza maxima	Tombusviridae	8	41,51-61,70
1E-058	Ehrharta calycina	Partitiviridae	1	48,57
1E-068	Avena byzantina	Geminiviridae	1	44,59
1E-070	Bromus diandrus	Nanoviridae	27	47,62-62,07
1E-074	Lolium perenne	Totiviridae	9	51,85-59,26
1E-074	Lolium perenne	Secoviridae	1	40,74
1E-074	Lolium perenne	Virgaviridae	1	29,52
1E-094	Triticum turgidum	Partitiviridae	7	45,45-49,57
1F-025	Willdenowia incurvata	Nanoviridae	3	38,10-41,03
1F-028	Bromus diandrus	Caulimoviridae	1	39,36
1F-068	Bromus diandrus	Partitiviridae	4	32,99-48,53
1F-069	Avena byzantina	Geminiviridae	6	37,04-38,18
1F-069	Avena byzantina	Nanoviridae	3	42,25-45
1F-069	Avena byzantina	Tombusviridae	16	34,48-59,18
1F-073	Lebeckia sepiaria	Alphaflexiviridae	3	51,43-61,22
1F-099	Avena byzantina	Tymoviridae	8	29,73
1G-061	Panicum sp	Partitiviridae	7	37,74-41,07
1H-028	Bromus diandrus	Secoviridae	2	42,86
1J-021	Inconnu	Alphaflexiviridae	10	31,90-38,46
1J-072	Euphorbia caput-medusae	Partitiviridae	17	38,78-59,65
1J-075	Ehrharta calycina	Partitiviridae	6	56,82-65,75
1K-051	Panicum sp	Partitiviridae	3	48
1K-090	Hordeum murinum	Geminiviridae	2	43,01-61,29
F2012				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2-08-D	Stipagrostis sp	Caulimoviridae	2	44,1-41,27
2-10-A	Inconnu	Nanoviridae	4	50-60
2-10-F	Stipagrostis	Caulimoviridae	63	32,20-60,98
2-12-A	Galenia sp	dsRNA unclassified	1	60,61
2-12-B	Lolium sp	Partitiviridae	3	58,14
2-14-G	Asparagus suaveolens	Closteroviridae	12	28,44-61,02
2-19-C	Muraltea spinosa	Caulimoviridae	24	36,51-58,54
2-21-D	Adenogramma glomerata	Partitiviridae	33	39,47-48,10
2-22-H	Salvia africana-coerulea	Caulimoviridae	3	33,63-35,96
2-23-G	Passurina sp.	Geminiviridae	1	52

F2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2-25-C	Nemesia versicolor	Totiviridae	2	35,8
2-26-B	Salvia africana	Caulimoviridae	3	49,3-49,32
2-26-E	Stipagrostis sp	Caulimoviridae	1	37,7
2-28-E	Inconnu	Geminiviridae	5	50
2-28-G	Lolium perenne	Geminiviridae	2	40,58-44,44
2-28-G	Inconnu	Nanoviridae	1	38,89
2-29-B	Willdenowia sp	Caulimoviridae	1	41,18
2-31-D	Asparagus suaveolens	Geminiviridae	2	45,83
2-31-G	Searsia glauca	Tymoviridae	3	46,67
2-32-D	Adenogramma glomerata	Nanoviridae	3	54,24
2-33-E	Adenogramma glomerata	Caulimoviridae	1	40,74
2-33-E	Inconnu	dsRNA unclassified	6	56
2-34-A	Helichrysum revolutum	Caulimoviridae	1	40
2-40-D	Nemesia versicolor	Caulimoviridae	8	33,71-64,29
2-41-A	Adenogramma glomerata	Partitiviridae	47	44,44-61,76
2-41-A	Inconnu	dsRNA unclassified	3	55,56-69,23
2-44-E	Inconnu	Geminiviridae	1	50,91
2-45-D	Albuca sp.	Endornaviridae	2	43,10-53,85
2-47-H	Eragrostis capensis	Caulimoviridae	1	35,63
2-48-G	Inconnu	Geminiviridae	11	34,92-46,67
2-50-E	Eragrostis capensis	Nanoviridae	2	46,15-48,08
2-51-I	Cysticapnos vesicarius	Caulimoviridae	1	33,72
2-53-F	Panicum ecklonii	Tombusviridae	4	40,54-41,18
2-55-B	Stipagrostis sp	Geminiviridae	2	26,32-39,77
2-55-B	Inconnu	Caulimoviridae	1	46,67
2-55-B	Inconnu	Totiviridae	10	50-56,86
2-57-C	Arctopus echinatus	Geminiviridae	1	46,15
2-58-E	Wahlenbergia capensis	Caulimoviridae	2	37,8
2-62-B	Inconnu	Betaflexiviridae	1	54,02
2-62-C	Felicia sp.	Caulimoviridae	7	33,33-53,49
2-66-G	Trilobium uniolae	Geminiviridae	15	33,68-50
2-67-A	Trilobium uniolae	Caulimoviridae	1	57,89
2-71-A	Panicum ecklonii	Geminiviridae	2	32,23-40,71
2-71-F	Salvia africana-coerulea	Geminiviridae	1	38,96
2-72-B	Inconnu	Closteroviridae	2	34,38-48,24
2-72-B	Inconnu	dsRNA virus	1	63,64
2-75-B	Rumex lativalvis	Caulimoviridae	1	45,88
2-90-I	Inconnu	Geminiviridae	3	52,83-61,11
2-90-J	Lolium perenne	Nanoviridae	3	46,94
2-96-C	Avena sativa	Geminiviridae	11	35,29-62,16
2-96-E	Hordeum vulgare	Geminiviridae	4	47,83-62,16
2-69-E	Avena sativa	Tymoviridae	1	48,44
2-97-B	Avena sativa	Geminiviridae	1	56,34
2-70-D	Inconnu	Partitiviridae	1	45,55
2-70-D	Inconnu	Geminiviridae	1	47,37

F2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2-69-A	Lolim perenne	Geminiviridae	6	55,56-60,61
2-89-E	Tetragonia fruticosa	Secoviridae	2	45,16
2-94-F	Bromus diandrus	Geminiviridae	12	45,45-52,63
2-99-E	Hordeum vulgare	Partitiviridae	24	52,31-64,15
2-74-A	Cynodon sp	Partitiviridae	2	46,30-59,38
2-77-B	Panicum ecklonii	Caulimoviridae	22	40,48-58,97
2-94-B	Phalaris minor	Caulimoviridae	1	38,46
2-100-I	Exomis sp	Endornaviridae	6	46,15-69,77
2-37-E	Limeum africanum	Nanoviridae	1	37,97
2-70-G	Bromus diandrus	Geminiviridae	9	32,91-48,57
2-70-G	Bromus diandrus	Caulimoviridae	1	52,54
2-95-F	Eragrostis capensis	Geminiviridae	1	50,88
2-89-C	Brassica oleraceae	Caulimoviridae	19	33,33-42,55
2-77-E	Avena sativa	Caulimoviridae	1	42.17
2-37-A	Lolium perenne	Partitiviridae	1	48,08
2-100-K	Lolium perenne	Caulimoviridae	6	38,64-41,79
2-86-F	Lolium perenne	Partitiviridae	5	49,02-65
C2010				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
01-1D	Plantago coronopus	Geminiviridae	11	48,33-52,78
02-1B	Trifolium nigrescens	Partitiviridae	4	36,78-40,15
03-1B	Halimione portulacoides	Geminiviridae	1	58,06
03-1C	Bromus hordeaceus	Geminiviridae	5	35,43-39,77
03-1F	Bellis annua	Nanoviridae	2	39,77-47,69
03-1G	Polypogon maritimus	Nanoviridae	2	28,57-31,08
03-1G	Polypogon maritimus	Geminiviridae	1	37,65
04-1B	Puccinellia festuciformis	Caulimoviridae	2	29,94-49,62
04-1B	Puccinellia festuciformis	Nanoviridae	47	35,05-66,67
11-1B	Polypogon maritimus	Nanoviridae	20	34,29-61,29
12-1D	Ranunculus peltatus	Geminiviridae	19	29,17-60,87
13-1E	Bromus madritensis	Tymoviridae	1	30,77
13-1E	Bromus madritensis	Geminiviridae	6	35,29-41,54
13-1F	Dactylis hispanica	Geminiviridae	2	28,87-32,94
13-1F	Dactylis hispanica	Nanoviridae	2	33,94-45,45
13-1F	Dactylis hispanica	Tombusviridae	46	24,22-30
14-1B	Suaeda vera	Nanoviridae	82	41,33-50
14-1C	Halimione portulacoides	Geminiviridae	10	38,89-57,89
14-1E	Bellis annua	Nanoviridae	2	40-42,03
14-1G	Bromus madritensis	Geminiviridae	9	34,38-35,16
21-1A	Carex cuprina	Partitiviridae	13	40,29-59,68
23-1A	Hordeum marinum	Geminiviridae	3	36,56-57,58
23-1C	Inconnu	Geminiviridae	1	54,17
24-1G	Festuca arundinacea	Partitiviridae	2	58,9-59,09

C2010 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
24-1G	Festuca arundinacea	ds RNA virus	7	35,38-54,17
25-1A	Inconnu	Tombusviridae	1	29,41
25-1A	Inconnu	Caulimoviridae	2	40,23
25-1D	Phragmites australis	Geminiviridae	25	30,05-68,18
31-1A	Festuca arundinacea	Geminiviridae	1	30,77
31-1A	Festuca arundinacea	ds RNA virus	2	53,52-57,75
31-1B	Plantago major	Geminiviridae	2	40,70-47,37
31-1C	Ranunculus sardous	Geminiviridae	1	33,78
31-1C	Ranunculus sardous	Nanoviridae	14	32,09-55
31-1E	Inconnu	Secoviridae	9	46,07-50,7
32-1A	Inconnu	Partitiviridae	11	57,14-61,26
33-1B	Lotus corniculatus	Partitiviridae	8	51,61-66,91
33-1D	Hordeum sp.	Geminiviridae	6	33,01-61,54
34-1A	Lotus glaber	Benyvirus	20	25,93-40,85
34-1A	Lotus glaber	Closteroviridae	102	20,12-63,85
34-1A	Lotus glaber	Secoviridae	112	28,57-54,76
34-1D	Trifolium repens	Luteoviridae	1	61,11
41-1E	Puccinellia festuciformis	ds RNA virus	18	25,52-60,71
43-1D	Trifolium repens	Endornaviridae	1	35,29
44-1A	Festuca arundinacea	Nanoviridae	8	38,03
44-1D	Trifolium repens	Secoviridae	168	27,78-63,,27
44-1F	Lotus glaber	Partitiviridae	2	56,25-71,88
45-1C	Hordeum marinum	Endornaviridae	1	63,46
45-1D	Inconnu	Nanoviridae	1	44,12
51-1A	Trifolium resupinatum	Partitiviridae	12	49,06-68,75
51-1C	Medicago polymorpha	Geminiviridae	4	37,04-59,46
53-1D	Puccinellia festuciformis	ds RNA virus	2	32,29-33,33
55-1D	Trifolium resupinatum	Geminiviridae	5	36,05-51,16
61-1G	Bromus hordeaceus	Partitiviridae	16	43,88-61,90
61-1G	Bromus hordeaceus	Geminiviridae	31	38,46-54,55
62-1B	Aeluropus littoralis	Geminiviridae	36	35,29-65,52
64-1C	Carex divisa	Nanoviridae	16	36,89-54,17
65-1C	Hordeum marinum	Geminiviridae	2	54,41-60,47
65-1C	Hordeum marinum	Nanoviridae	79	38,57-62,86
71-1A	Campanula rapunculis	Geminiviridae	2	48,98
72-1A	Puccinellia sp.	ds RNA virus	2	50-58,76
72-1A	Puccinellia sp.	Tombusviridae	2	30,50-31,61
72-1C	Lolium rigidum	Geminiviridae	1	34,48
73-1A	Galium aparinella	Tombusviridae	24	29-34,29
74-1A	Puccinellia festuciformis	Tombusviridae	3	33,33-36,67
74-1A	Puccinellia festuciformis	Nanoviridae	1	45,83
74-1A	Puccinellia festuciformis	Geminiviridae	1	48,72
74-1A	Puccinellia festuciformis	Partitiviridae	3	23,36-25
74-1A	Puccinellia festuciformis	ds RNA virus	47	25,9-66,09

C2010 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
75-1C	Festuca arundinacea	Caulimoviridae	8	35,66-42,47
75-1C	Festuca arundinacea	Potyviridae	21	33,66-51,32
82-1A	Ranunculus peltatus	Endornaviridae	1	31,48
82-1A	Ranunculus peltatus	Nanoviridae	1	45,45
82-1A	Ranunculus peltatus	Geminiviridae	15	37,93-50
82-1D	Hordeum sp.	Geminiviridae	1	35,37
83-1C	Bromus hordeaceus	Tombusviridae	3	32,35-36,26
83-1C	Bromus hordeaceus	Nanoviridae	1	35,62
86-1A	Hordeum marinum	Geminiviridae	6	38,89-47,92
87-1B	Elytrigia campestre	Geminiviridae	1	50,94
92-1B	Medicago tribuloides	Nanoviridae	1	35,9
96-1A	Limonium narbonense	Tombusviridae	22	30,77-37,74
05-1B	Oryza sativa	Geminiviridae	1	43,18
07-1B	Oryza sativa	Nanoviridae	2	48,84
10-1A	Triticum sp.	Virgaviridae	16	30,77-32,48
17-1E	Inconnu	Partitiviridae	3	45,24-47,56
20-1D	Inconnu	Tombusviridae	2	36,89-38,53
20-1D	Inconnu	Geminiviridae	2	23,57-44,23
20-1D	Inconnu	Nanoviridae	5	35,77-68,29
20-1E	Trifolium repens	Flexiviridae	1	67,24
20-1E	Trifolium repens	Betaflexiviridae	1	61,34
29-1B	Chenopodium sp.	Partitiviridae	185	28,95-71,43
35-1D	Inconnu	Geminiviridae	4	40,28-64,10
36-1D	Juncus gerardii	Nanoviridae	3	41,79-45,28
38-1C	Oryza sativa	Endornaviridae	67	57,83
39-1B	Phragmites australis	Geminiviridae	79	31,67-62,50
47-1A	Lotus corniculatus	Betaflexiviridae	96	42,5-93,18
47-1C	Trifolium repens	Closteroviridae	14	30,77-57,14
48-1B	Medicago sativa	ds RNA virus	1	56,82
50-1B	Oryza sativa	Nanoviridae	19	33,33-61,22
50-1C	Oryza sativa	Geminiviridae	1	48,61
60-1A	Oryza sativa	Geminiviridae	1	35,85
60-1D	Oryza sativa	Geminiviridae	2	48,48
69-1C	Oryza sativa	Nanoviridae	2	45,71
78-1A	Spegularia salina	Endornaviridae	3	44,48-55,71
78-1D	Franckenia pulverulenta	Nanoviridae	1	56,71
79-1B	Echinochloa crus-galli	Geminiviridae	5	30,69
80-1C	Phragmites australis	ds RNA virus	40	50-67,74
88-1D	Phragmites australis	Alphaflexiviridae	1	54,9
C2012				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-01-B	Arthrocnemum macrostachyum	Partitiviridae	9	35,06-67,14
2012-02-C	Halimione portulacoides	Caulimovidae	1	30,68
2012-02-C	Halimione portulacoides	Endonaviridae	262	22,22-67,65

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-03-A	Halimione portulacoides	Nanoviridae	1	48
2012-03-A	Halimione portulacoides	Endonaviridae	9	28,18-56,86
2012-03-B	Bromus madritensis	Geminiviridae	4	40,26-45,90
2012-04-E	Halimione portulacoides	Nanoviridae	10	33,61-55,10
2012-06-C	Arthrocnemum macrostachyum	Partitiviridae	18	38,71-64,44
2012-11-B	Hordeum marinum	Geminiviridae	1	40
2012-11-C	Halimione portulacoides	Caulimovidae	2	41,6
2012-11-E	Elytrigia atherica	Totiviridae	7	35,71-60
2012-11-E	Elytrigia atherica	Geminiviridae	107	26,52-85
2012-13-B	Halimione portulacoides	Endornaviridae	7	32,89-58,27
2012-14-A	Halimione portulacoides	Nanoviridae	7	35,79-60
2012-15-B	Atriplex prostrata	Partitiviridae	26	32,50-49,21
2012-92-D	Arthrocnemum macrostachyum	Luteoviridae	1	63,64
2012-15-F	Hordeum marinum	Caulimoviridae	1	46,55
2012-15-F	Hordeum marinum	Geminiviridae	3	36,14-37,36
2012-15-H	Puccinellia sp.	Partitiviridae	15	33,88-48,05
2012-21-A	Festuca arundinacea	dsRNA Virus	16	48,10-64,38
2012-21-E	Carex cuprina	Partitiviridae	1	49,09
2012-22-F	Iris pseudacorus	dsRNA Virus	1	45,28
2012-23-E	Hordeum murinum	Tymoviridae	2	60,47-61,36
2012-23-J	Plantago coronopus	Geminiviridae	2	45,45
2012-25-B	Polypogon maritimus	Nanoviridae	2	36,17-47,31
2012-25-D	Spergularia sp.	dsRNA Virus	2	50,91-64,86
2012-25-D	Spergularia sp.	Tombusviridae	4	32,12-37,21
2012-25-E	Juncus gerardii	Nanoviridae	2	44,74-48,98
2012-31-A	Lythrum salicaria	Caulimovidae	3	38,89-40,85
2012-31-D	Inconnu	Geminiviridae	11	30,26-47,17
2012-31-E	Galium debile	Geminiviridae	1	46,43
2012-31-G	Festuca arundinacea	dsRNA Virus	4	59,09
2012-31-G	Festuca arundinacea	Partitiviridae	3	55-68,57
2012-31-G	Festuca arundinacea	Geminiviridae	5	40-53,85
2012-31-H	Festuca arundinacea	Geminiviridae	10	31,08-41,18
2012-31-L	Trifolium pratense	Totiviridae	1	53,54
2012-31-L	Trifolium pratense	Bromoviridae	4	34,78-46,15
2012-31-L	Trifolium pratense	Closteroviridae	21	34,78-45,28
2012-31-L	Trifolium pratense	Virgaviridae	4	37,5
2012-31-L	Trifolium pratense	Geminiviridae	4	46,43
2012-32-F	Ranunculus bulbosus	Geminiviridae	19	32,43-40,2
2012-32-J	Trifolium resupinatum	dsRNA Virus	2	54,95-60,47
2012-32-J	Trifolium resupinatum	Totiviridae	2	28,99-65,38
2012-33-A	Lotus sp.	Totiviridae	10	36,9-38,82
2012-33-D	Hordeum marinum	Geminiviridae	2	47,46-49,02
2012-33-D	Hordeum marinum	Nanoviridae	11	55,77-63,16
2012-33-E	Festuca arundinacea	Partitiviridae	78	37,29-69,66

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-34-F	Festuca arundinacea	dsRNA Virus	4	56,14-67,36
2012-41-E	Asparagus scaber	dsRNA Virus	1	67,54
2012-41-E	Asparagus scaber	Partitiviridae	1	69,74
2012-41-N	Puccinellia festuciformis	dsRNA Virus	36	30,21-59,32
2012-42-A	Carex cuprina	Tymoviridae	13	39,56-53,85
2012-42-E	Picris echioides	Caulimovidae	1	40,85
2012-42-G	Lotus corniculatus	Geminiviridae	3	37,80-49,23
2012-42-H	Festuca arundinacea	dsRNA Virus	2	40,28-58,33
2012-44-A	Cynodon dactylon	Geminiviridae	1	33,33
2012-44-A	Cynodon dactylon	Partitiviridae	7	52,17-62,5
2012-44-A	Cynodon dactylon	dsRNA Virus	26	43,96-60,42
2012-44-D	Agrostis stolonifera	Alphaflexiviridae	1	34,09
2012-44-E	Festuca arundinacea	dsRNA Virus	3	46,48-56,25
2012-44-F	Medicago sativa	Partitiviridae	37	37,21-63,64
2012-45-D	Hordeum marinum	Geminiviridae	4	44,86
2012-45-I	Elytrigia atherica	Geminiviridae	21	40,48-55,88
2012-45-K	Aster tripolium	Nanoviridae	1	40,68
2012-51-Bbis	Halimione portulacoides	Partitiviridae	4	51,11-63,64
2012-51-D	Halimione portulacoides	Caulimovidae	1	52,63
2012-51-E	Bromus madritensis	Geminiviridae	1	56,41
2012-51-E	Bromus madritensis	Caulimovidae	9	29,55-34
2012-52-B	Dorycnium	Partitiviridae	1	34,38
2012-52-D	Elytrigia atherica	Nanoviridae	2	49,44
2012-52-E	Avena barbata	Geminiviridae	2	45,65
2012-52-F	Halimione portulacoides	Partitiviridae	1	68
2012-53-A	Elytrigia elongata	Caulimovidae	1	49,21
2012-55-F	Elytrigia atherica	Geminiviridae	17	36,47-63,33
2012-58-A	Festuca arundinacea	Partitiviridae	2	58,18
2012-58-A	Festuca arundinacea	Geminiviridae	4	26,67-32,91
2012-58-A	Festuca arundinacea	Virgaviridae	15	50-97,06
2012-58-D	Oxalis sp.	Partitiviridae	1	50
2012-62-B	Hordeum marinum	Geminiviridae	2	33,33-38,89
2012-62-D	Ranunculus parviflora	Partitiviridae	2	43,69-56,86
2012-65-C	Arthrocnemum macrostachyum	dsRNA Virus	1	62,79
2012-74-F	Sarcocornia fruticosa	ssRNA + virus	1	46,3
2012-74-F	Sarcocornia fruticosa	Caulimovidae	5	37,17-59,55
2012-74-H	Festuca glyceria	dsRNA Virus	53	23,02-61,36
2012-75-B	Elytrigia atherica	Tombusviridae	2	51,06-54,55
2012-75-D	Carex cuprina	Alphaflexiviridae	6	40,32-52,94
2012-82-A	Arthrocnemum macrostachyum	Partitiviridae	19	38,26-65,33
2012-43-K	Trifolium repens	Luteoviridae	2	56,14-67,92
2012-86-A	Spergularia sp.	Nanoviridae	1	49,18
2012-87-C	Puccinellia festuciformis	Partitiviridae	20	30,51-52,24
2012-87-E	Limonium narbonense	Totiviridae	3	38,20-68,42

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-92-F	Elytrigia sp.	Tymoviridae	10	52,50-55
2012-96-D	Arthrocnemum macrostachyum	Partitiviridae	12	43,75-69,57
2012-96-D	Arthrocnemum macrostachyum	Tombusviridae	1	46,91
2012-05-C	Oryza sativa	Tombusviridae	1	52,86
2012-07-A	Triticum turgidum	Totiviridae	1	43,86
2012-09-A	Avena sp.	dsRNA virus	1	49,33
2012-09-C	Bidens sp.	Partitiviridae	2	51,22
2012-100-A	Sonchus maritimus	Caulimoviridae	1	67,65
2012-16-D	Festuca arundinacea	Partitiviridae	4	46,81
2012-16-D	Festuca arundinacea	dsRNA virus	1	50
2012-56-F	Inconnu	Luteoviridae	1	65,88
2012-19-E	Phragmites australis	Tombusviridae	1	43,75
2012-26-A	Oryza sativa	dsRNA virus	1	58,23
2012-27-A	Medicago sativa	Partitiviridae	32	43,33-60,87
2012-27-A	Medicago sativa	Tymoviridae	2	42,03-63,16
2012-27-G	Plantago coronopus	Geminiviridae	1	43,64
2012-28-B	Trifolium angustifolium	Partitiviridae	14	40,66-59,30
2012-30-E	Euphorbia sp.	Tombusviridae	1	66,67
2012-05-C	Oryza sativa	Luteoviridae	1	54,55
2012-36-A	Festuca arundinacea	Partitiviridae	3	48,94-54,84
2012-36-E	Chrysanthemum sp.	Partitiviridae	1	58,33
2012-36-E	Chrysanthemum sp.	Geminiviridae	3	43,75
2012-38-C	Triticum turgidum	Partitiviridae	2	39,53-40,70
2012-38-C	Triticum turgidum	Reoviridae	1	37,6
2012-84-E	Oryza sativa	Luteoviridae	1	49,21
2012-38-F	Oxalis sp.	dsRNA virus	7	45,16-63,46
2012-35-D	Phragmites australis	Luteoviridae	1	44,44
2012-39-B	Oryza sativa	Endornaviridae	218	56,36
2012-46-B	Oryza sativa	Geminiviridae	3	52,13-55,38
2012-47-D	Trifolium sp.	Tombusviridae	1	45,61
2012-48-C	Medicago sativa	Tymoviridae	7	39,47-55,56
2012-48-D	Trifolium sp.	Geminiviridae	2	42,86-44,44
2012-48-D	Trifolium sp.	Dicistroviridae	1	54,95
2012-48-E	Medicago sativa	dsRNA virus	6	50-65,22
2012-48-E	Medicago sativa	Tymoviridae	3	45,28-52,83
2012-48-E	Medicago sativa	Partitiviridae	4	51,90-66,67
2012-54-D	Festuca arundinacea	Partitiviridae	2	37,31-68,63
2012-54-F	Medicago sativa	Partitiviridae	191	37,4-89,58
2012-56-E	Inconnu	Totiviridae	9	45,63-52,83
2012-67-A	Phragmites australis	Geminiviridae	7	30,68-47,44
2012-67-B	Bolboschoenus maritimus	Geminiviridae	19	32,56-48,94
2012-67-C	Oryza sativa	Geminiviridae	2	30,70-47,42
2012-67-D	Oryza sativa	Geminiviridae	58	28,57-65,52
2012-68-A	Medicago sativa	Nanoviridae	2	33,33-35,82

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-68-A	Medicago sativa	Geminiviridae	1	29,17
2012-68-A	Medicago sativa	Partitiviridae	6	28,57-65,52
2012-68-B	Cynodon dactylon	Nanoviridae	1	33,78
2012-68-B	Cynodon dactylon	Geminiviridae	3	33,33-41,46
2012-68-B	Cynodon dactylon	Partitiviridae	6	30,23-54,72
2012-76-D	Sorghum bicolor	dsRNA virus	6	47,31-57,41
2012-78-G	Puccinellia fasciculata	Tombusviridae	87	34,67-60
2012-79-A	Cynodon dactylon	Geminiviridae	3	28,42-29,75
2012-79-A	Cynodon dactylon	Nanoviridae	27	44,12-50
2012-79-E	Oryza sativa	Nanoviridae	1	50,98
2012-80-E	Oryza sativa	Geminiviridae	1	45,88
2012-80-F	Alisma plantago	Geminiviridae	1	51,43
2012-84-D	Oryza sativa	Geminiviridae	2	30,53
2012-84-D	Oryza sativa	Endornaviridae	1560	44,19
2012-84-E	Oryza sativa	Geminiviridae	8	28,85-50
2012-89-B	Festuca arundinacea	Partitiviridae	1	41,67
2012-89-D	Bolboschoenus maritimus	Umbravirus	2	48,03-48,25
2012-89-D	Bolboschoenus maritimus	Tombusviridae	1	46,91
2012-90-B	Oryza sativa	Geminiviridae	1	50,7
2012-90-E	Bolboschoenus maritimus	dsRNA virus	1	56,16
2012-94-G	Cynodon dactylon	Nanoviridae	17	36,26-42,42
2012-97-A	Oryza sativa	dsRNA virus	1	58,75

Annexe 6 :

Tableau répertoriant les nouveaux variants/isolats viraux potentiels découverts via nos travaux de géo-métagénomique.

F2010				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
1A-027	Bromus diandrus	Partitiviridae	47	32,52-77,78
1A-053	Passerina corymbosa	Partitiviridae	7	53,73-79,17
1A-085	Briza maxima	Tombusviridae	8	42,42-82,86
1B-072	Rumex lativalvis	Tombusviridae	136	37,37-80,56
1B-094	Triticum turgidum	Luteoviridae	145	60-100
1C-024	Indigofera heterophylla	Betaflexiviridae	697	34,94-79,17
1C-093	Phalaris minor	Partitiviridae	37	40,24-81,94
1D-012	Galenia africana	Partitiviridae	20	28,97-85,25
1D-012	Galenia africana	Secoviridae	105	35,38-79,27
1D-024	Dipogon lignosus	Potiviridae	216	35,29-93,33
1D-057	Clutia alaternoides	Tombusviridae	137	50-96,65
1D-073	Capnophyllum africana	Partitiviridae	102	33,94-82,88
1D-073	Capnophyllum africana	Totiviridae	13	37,88-86-36
1E-005	Capnophyllum africana	Partitiviridae	2	59,66-75,86
1E-023	Dipogon lignosus	Potyviridae	1019	35-97,30
1E-068	Avena byzantina	Potyviridae	3	69,32-95,74
1E-074	Lolium perenne	Partitiviridae	228	28,24-94,74
1E-074	Lolium perenne	Luteoviridae	52	55,80-100
1E-080	Asparagus rubicundus	Geminiviridae	35	43,42-77,27
1E-100	Cysticapnos vesicaria	Luteoviridae	16	58,82-100
1F-069	Avena byzantina	Partitiviridae	17	46,91-86,11
1F-073	Lebeckia sepiaria	Closteroviridae	26	43,42-77,94
1G-010	Staberoha distachyos	classified Sobemovi	45	37,93-76,12
1G-028	Tulbaghia capensis	Partitiviridae	42	47,37-88,89
1G-078	Cynodon dactylon	Partitiviridae	16	44,83-81,65
1H-090	Exomis sp	Geminiviridae	249	43,43-100
1I-031	Dipogon lignosus	Potyviridae	329	31,52-94,92
1I-083	Inconnu	Partitiviridae	65	34,67-81,82
F2012				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2-09-E	Anthospermum aethiopicum	Caulimoviridae	2	65-81,94
2-12-A	Inconnu	Secoviridae	6	36,15-75,86
2-12-F	Limeum africanum	Geminiviridae	7	58,33-88,46
2-14-B	Inconnu	Geminiviridae	12	43,10-78,57
2-32-A	Aspalathus sp.	Potyviridae	7	50-88
2-36-E	Pelagornium	Caulimoviridae	2	67,50-70
2-43-B	Pentaschistis	Partitiviridae	4	50-72,60
2-62-B	Inconnu	Partitiviridae	103	52,70-100
2-71-C	Capnophyllum africanum	Partitiviridae	57	29,73-95,83
2-71-C	Inconnu	Totiviridae	57	35,29-84,38
2-83-C	Inconnu	Potyviridae	55	67,50-100
2-90-C	Exomis sp	Geminiviridae	194	46,15-98-33

C2010				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
14-1C	Halimione portulacoides	Nanoviridae	232	30,19-70,83
22-1A	Festuca arundinacea	ds RNA virus	35	45,12-70,37
32-1A	Inconnu	Totiviridae	3	69,05-72,06
33-1B	Lotus corniculatus	ds RNA virus	45	45,05-73,58
33-1C	Festuca arundinacea	Virgaviridae	18	54,12-100
34-1A	Lotus glaber	Betaflexiviridae	207	31,30-89,58
34-1D	Trifolium repens	Partitiviridae	229	45,97-100
34-1E	Festuca arundinacea	Virgaviridae	48	40,26-95,92
44-1E	Medicago sativa	Geminiviridae	48	45,78-82,93
44-1F	Lotus glaber	Partitiviridae	2	56,25-71,88
44-1F	Lotus glaber	ds RNA virus	9	46,88-73,81
53-1A	Asparagus acutifolius	Partitiviridae	652	29,55-80,65
61-1B	Crepis virens	Caulimoviridae	25	53,23-100
61-1E	Asparagus acutifolius	Partitiviridae	24	30,93-79,37
72-1C	Lolium rigidum	Partitiviridae	66	31,67-84,91
73-1D	Atriplex halimus	Endornaviridae	50	25-77,55
81-1A	Inconnu	Partitiviridae	53	38,83-85,71
10-1C	Triticum sp.	Luteoviridae	5	66,51-98,90
10-1E	Triticum sp.	Luteoviridae	10	67,39-100
17-1B	Euphorbia cypacicias	Totiviridae	2	64,86-76,47
18-1B	Trifolium repens	Potyviridae	73	52,34-100
19-1A	Oryza sativa	Endornaviridae	8	56,03-97,87
19-1D	Oryza sativa	Caulimoviridae	51	50,39-100
20-1E	Trifolium repens	Alphaflexiviridae	3975	38,71-100
27-1B	Hordeum sp.	Potyviridae	18	65,05-100
29-1B	Chenopodium sp.	Partitiviridae	185	28,95-71,43
30-1G	Vicia cracca	Nanoviridae	1408	33,81-97,06
36-1A	Festuca arundinacea	Virgaviridae	134	48,96-97,50
36-1A	Festuca arundinacea	Partitiviridae	1003	34,45-97,67
38-1D	Oryza sativa	Endornaviridae	2	55,36-93,94
46-1B	Oryza sativa	Endornaviridae	18	60,61-100
46-1C	Oryza sativa	Endornaviridae	23	61,54-100
47-1A	Lotus corniculatus	Luteoviridae	11	48,15-71,64
47-1A	Lotus corniculatus	Secoviridae	6	47,11-78,26
48-1B	Medicago sativa	Partitiviridae	21	46,08-78,15
48-1F	Melilotus alba	ds RNA virus	41	40,91-78,57
48-1F	Melilotus alba	Partitiviridae	44	25,53-77,42
49-1C	Triticum sp.	Luteoviridae	17	65,43-100
49-1DBIS	Solanum lycopersicum	ds RNA virus	69	43,59-100
60-1A	Oryza sativa	Endornaviridae	24	55,56-100
68-1B	Triticum sp.	Luteoviridae	75	47,42-100
C2012				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-03-C	Avena barbata	Potyviridae	8	57,30-98,41

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-03-D	Suaeda vera	Bromoviridae	4	63,46-85,19
2012-03-E	Sarcocornia fruticosa	Partitiviridae	2	68,63-100
2012-11-C	Halimione portulacoides	Potyviridae	57	44,12-98,11
2012-11-E	Elytrigia atherica	Geminiviridae	107	26,52-85
2012-12-A	Arthrocnemum macrostachyum	Partitiviridae	124	33,33-75
2012-13-F	Juncus maritimus	Geminiviridae	168	38,71-91,89
2012-82-C	Aeluropus littoralis	Luteoviridae	93	41,07-94,55
2012-21-A	Festuca arundinacea	Partitiviridae	263	37,24-96,15
2012-25-D	Spergularia sp.	Virgaviridae	7	57,41-88,68
2012-31-L	Trifolium pratense	Partitiviridae	5	54,17-77,89
2012-32-J	Trifolium resupinatum	Partitiviridae	6	38,71-72,44
2012-34-F	Festuca arundinacea	Partitiviridae	22	45,74-84,95
2012-41-N	Puccinellia festuciformis	Partitiviridae	301	32,47-100
2012-42-H	Festuca arundinacea	Partitiviridae	247	42,22-100
2012-42-H	Festuca arundinacea	Virgaviridae	424	46,27-100
2012-43-E	Hordeum secale	Partitiviridae	3	40,96-70,53
2012-43-E	Hordeum secale	dsRNA Virus	36	36,11-70
2012-43-E	Hordeum secale	Luteoviridae	9	65,88-100
2012-44-B	Festuca arundinacea	dsRNA Virus	8	61,76-79,31
2012-44-B	Festuca arundinacea	Virgaviridae	16	57,78-85
2012-44-C	Carex cuprina	Caulimovidae	31	56,90-100
2012-44-E	Festuca arundinacea	Partitiviridae	85	44,44-93,62
2012-44-E	Festuca arundinacea	Virgaviridae	102	41,38-100
2012-44-F	Medicago sativa	dsRNA Virus	42	30,29-70,15
2012-45-I	Elytrigia atherica	Partitiviridae	4	67,86-80,33
2012-51-A	Sarcocornia fruticosa	Potyviridae	99	51,11-98,25
2012-58-A	Festuca arundinacea	Virgaviridae	15	50-97,06
2012-58-C	Circium arvense	Potyviridae	24	60-83,33
2012-58-G	Verbena officinalis	Closteroviridae	83	37,27-83,33
2012-61-A	Limonium narbonense	Geminiviridae	30	48,75-93,06
2012-65-C	Arthrocnemum macrostachyum	Potyviridae	21	68,18-96,30
2012-75-B	Elytrigia atherica	Umbravirus	93	36,36-78,79
2012-75-F	Brachypodium pheonicoides	Geminiviridae	106	35-70
2012-75-F	Brachypodium pheonicoides	Nanoviridae	673	30,25-84,62
2012-75-F	Brachypodium pheonicoides	Potyviridae	28	43,75-75,50
2012-41-N	Puccinellia festuciformis	Luteoviridae	19	65,45-97,47
2012-87-E	Limonium narbonense	Partitiviridae	25	32,94-81,51
2012-92-B	Halimione portulacoides	Endonaviridae	181	25,86-71,01
2012-05-B	Oryza sativa	Endornaviridae	279	45,68-100
2012-05-D	Oryza sativa	Endornaviridae	199	51,94-100
2012-05-E	Festuca arundinacea	Virgaviridae	27	57,69-95,35
2012-60-F	Bolboschoenus maritimus	Luteoviridae	42	63,16-90,48
2012-08-A	Oryza sativa	Endornaviridae	14	67,11-98,51

C2012 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-08-E	Oryza sativa	Endornaviridae	13	42-100
2012-09-A	Avena sp.	Partitiviridae	103	30,51-79,17
2012-09-A	Avena sp.	Virgaviridae	122	50-100
2012-100-G	Medicago lupulina	Potyviridae	18	41,74-92,16
2012-28-B	Trifolium angustifolium	Luteoviridae	3	42,22-98,56
2012-48-D	Trifolium sp.	Luteoviridae	76	64,79-100
2012-16-A	Oryza sativa	Endornaviridae	152	56,41-100
2012-16-C	Oryza sativa	Endornaviridae	9	65,67-100
2012-17-A	Avena barbata	Potyviridae	173	45,83-100
2012-17-F	Foeniculum vulgare	Partitiviridae	94	47,15-94,12
2012-17-G	Pinus halepensis	Totiviridae	59	30,95-78,26
2012-18-A	Arundo sp.	Caulimoviridae	36	49,37-85,57
2012-18-B	Oryza sativa	Endornaviridae	60	67,74-100
2012-18-E	Oryza sativa	Endornaviridae	8	52,38-100
2012-20-B	Plantago sp.	Tombusviridae	1372	35,06-91,30
2012-40-C	Echinochloa sp.	Luteoviridae	52	52,11-90,48
2012-26-A	Oryza sativa	Endornaviridae	57	49,37-100
2012-26-B	Oryza sativa	Endornaviridae	85	54,05-100
2012-27-A	Medicago sativa	dsRNA virus	43	27,73-80
2012-27-A	Medicago sativa	Nanoviridae	221	32,41-82,61
2012-60-B	Echinochloa sp.	Luteoviridae	121	47,37-96,30
2012-27-C	Atriplex prostrata	Partitiviridae	5	37,66-74,44
2012-27-C	Atriplex prostrata	Sobemovirus	1446	37,61-100
2012-28-B	Trifolium angustifolium	dsRNA virus	41	39,29-75
2012-10-A	Inconnu	Luteoviridae	1233	45,3-100
2012-30-B	Verbascum sinuatum	Partitiviridae	3	69,23-79,69
2012-38-G	Inconnu	Luteoviridae	11	57,80-98,80
2012-35-C	Oryza sativa	Endornaviridae	10	48,05-100
2012-36-A	Festuca arundinacea	dsRNA virus	99	50,88-71,43
2012-36-C	Festuca arundinacea	Partitiviridae	36	45,61-87,39
2012-37-D	Oryza sativa	Endornaviridae	96	47,46-100
2012-37-E	Oryza sativa	Endornaviridae	93	48,72-100
2012-38-F	Oxalis sp.	Partitiviridae	22	44,68-81,82
2012-39-A	Oryza sativa	Partitiviridae	2	45,36-72,13
2012-39-C	Avena barbata	Potyviridae	12	62,16-100
2012-39-F	Oryza sativa	Endornaviridae	226	48,19-100
2012-10-B	Triticum turgidum	Luteoviridae	241	51,52-100
2012-46-A	Oryza sativa	Endornaviridae	101	44,74-100
2012-46-B	Oryza sativa	Endornaviridae	25	53,76-100
2012-47-C	Trifolium resupinatum	Endornaviridae	39	54,67-100
2012-47-D	Trifolium sp.	Partitiviridae	13	32,14-70,83
2012-48-A	Medicago sativa	Geminiviridae	32	49,30-83,02
2012-48-B	Hordeum murinum	Potyviridae	16	58,97-100
2012-48-C	Medicago sativa	Partitiviridae	29	40,45-79,22

C2010 (suite)				
Echantillon	Plante	Famille virale	Nombre de reads	% identité (gamme)
2012-48-C	Medicago sativa	dsRNA virus	28	39,39-78,72
2012-10-F	Triticum turgidum	Luteoviridae	18	48,2-100
2012-54-D	Festuca arundinacea	Umbravirus	12	41,11-75
2012-54-D	Festuca arundinacea	Tombusviridae	17	43,33-83,33
2012-54-F	Medicago sativa	dsRNA virus	52	32-81,82
2012-54-F	Medicago sativa	Partitiviridae	191	37,4-89,58
2012-28-A	Triticum turgidum	Luteoviridae	81	56-100
2012-38-C	Triticum turgidum	Luteoviridae	385	55,10-100
2012-66-C	Oryza sativa	Endornaviridae	9	60,71-100
2012-66-D	Oryza sativa	Endornaviridae	135	53,57-100
2012-78-B	Lolium perenne	Secoviridae	3	65-87,10
2012-38-D	Triticum turgidum	Luteoviridae	70	44,25-100
2012-78-G	Puccinellia fasciculata	Umbravirus	33	39,80-70,59
2012-89-B	Festuca arundinacea	dsRNA virus	9	40-73,08
2012-90-B	Oryza sativa	Endornaviridae	12	69,84-100
2012-95-C	Oryza sativa	Endornaviridae	21	61,45-97,32
2012-97-A	Oryza sativa	Endornaviridae	101	64,37-100
2012-97-F	Oryza sativa	Endornaviridae	16	66,67-88,89
2012-99-D	Chenopodium album	dsRNA virus	324	51,02-100
2012-99-E	Solanum vilosum	dsRNA virus	119	53,25-100

Annexe 7 :

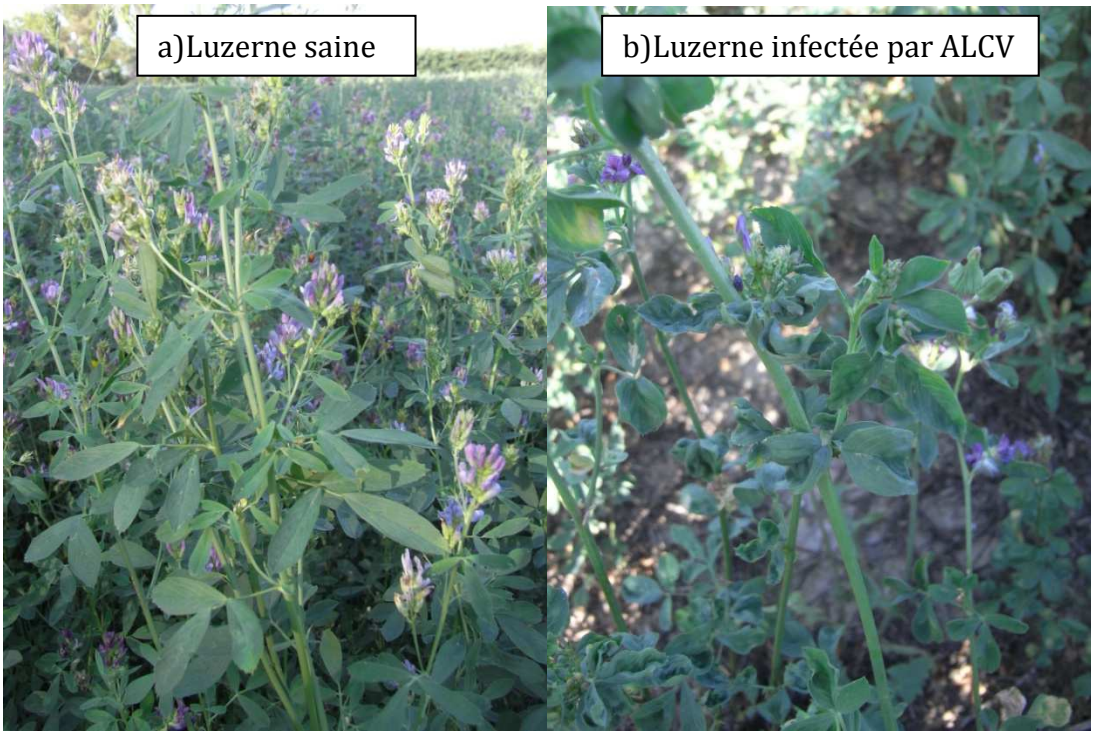
Séquence des couples d'amorces et conditions de PCR utilisées pour les amplifications relatives aux capulavirus.

Pour effectuer les amplifications nous avons utilisé soit la HotStar mastermix (Qiagen) soit la GoTaq Green Master Mix (Promega), les protocoles suivis concernant les milieux réactif et les cycles PCR sont ceux des fournisseurs.

Couples d'amorces	Séquence (5'-3')	Température d'hybridation (°C)	Température d'élongation (°C)	Temps d'élongation (s)
Dar-1981F	CCTCACTGAATCCACATCCA	60	68	180
Dar-1966R	CGAGGAATTCGGACTTGG			
Plantago-1F	CAGTCCACACTTCCGCAGTA	60	72	45
Plantago-1R	GGCCGAAACAAACCTCTACA			
Plantago-2F	CTGTGTTGCCCATGTGTAGG	60	72	45
Plantago-2R	AACCACACCACCCCAATATC			
Plantago-3F	AAGCCCTGTATCTGGGTTCA	60	72	45
Plantago-3R	ATGGAACGAAACACCTCCAC			
Plantago-4F	CTGTGTTGCCCATGTGTAGG	60	72	45
Plantago-4R	GAAATATGCCACGTCAACCA			
Plantago-5F	AATGGCAAAACGCTTCTAC	50	72	60
Plantago-5R	TAGCAATGATGGACTACG			
Plantago-6F	TGACGGACATGTTAATGTTCAA	50	72	60
Plantago-6R	TACTGCGGAAGTGTGGACTG			
Plantago-7F	CGGATGTTGGTGTTAAATCG	60	72	60
Plantago-7R	CTTCCCGTAACTGCTCGGTA			
Plantago-8F	CGAAGTTCCGTCATTCGATAA	50	72	60
Plantago-8R	CCAGACCAATCGTACAGAGC			
EcmLV-136-F	CGAAGAGGTCATTGGGACAT	60	72	40
EcmLV-730-R	CGGGTCTGGCTAAGAGAGTG			
EcmLV-1775-F	TTGAATTGCATGGGCACTTA	56	72	50
EcmLV-2433-R	GCCCTTTTGGTCATTTTGAA			
Luz-CP-F	TGGAATATTGTGCTGCTTGG	55	70	50
Luz-CP-R	ATTTTGGGACTTGTGCTCCA			

Annexe 8 :

a) Plant de luzerne sain
b) Plant de luzerne symptomatique infecté par *Alfafa leaf curl virus* (ALCV)


a)Luzerne saine
b)Luzerne infectée par ALCV

Annexe 9:

A novel Itera-like densovirus isolated by viral metagenomics from the sea barley *Hordeum marinum*.
S. François, P. Bernardo, D. Filloux, P. Roumagnac, D. Yaverkovski, R. Froissart & M. Ogliastro
Soumis à Genome Annoucements en Octobre 2014

A novel Itera-like densovirus isolated by viral metagenomics from the sea barley *Hordeum marinum.*

S. François[1#], P. Bernardo[2#], D. Filloux[2], P. Roumagnac[2], D. Yaverkovski[3], R. Froissart[2,4] & M. Ogliastro[1*]

[1] INRA, UMR 1333, DGIMI, F-34095 Montpellier, France.

[2] INRA-CIRAD-SupAgro, UMR 385, BGPI, Campus international de Baillarguet, 34398 Montpellier, France.

[3] Service écologie végétale, Fondation Tour du Valat, Le Sambuc, 13200 Arles, France.

[4] CNRS-IRD-UM1-UM2, UMR 5290, MIVEGEC, 911 avenue Agropolis, 34394 Montpellier France.

[*] corresponding authors : ogliastr@supagro.inra.fr

[#] these authors contributed equally to the work

ABSTRACT

Densoviruses (DVs) infect arthropods and belong to the *Parvoviridae* family. Here we report the complete coding sequence of a novel DV isolated from the plant *Hordeum marinum* (Poaceae) by viral metagenomics and confirmed re-amplification by PCR. Phylogenetic analyses showed that this novel DV is related to the genus *Iteradensovirus*.

GENOME ANNOUNCEMENT

Densoviruses (DVs) are small non-enveloped icosahedral viruses infecting arthropods, including pests and vectors for which they are considered as bio-control agents. They contain a single-strand linear DNA genome ranging from 4 to 6 kb ended by inverted terminal repeats (ITRs) [1]. Only 15 DV species are referenced in Genebank so far [2]; they display a large diversity of sequences, structures and organizations. Such diversity together with the diversity of their invertebrate hosts, suggest that DVs are largely unknown and ubiquitous in the environment. It is crucial to understand the densovirus diversity and prevalence for both fundamental and applied virology issues.

A novel densovirus was detected using VANA viral metagenomic approach [3], from sea barley (*Hordeum marinum*). To complement this genome we performed Rapid Amplification of cDNA Ends (RACE) (Roche) and products were

cloned in the pGEM-T Easy Vector (Promega) and sequenced. Sequences were assembled using Geneious 7.1.4 and compared to database sequences using BlastN, BlastP and tBlastX [4]. Results were considered as indicative of significant homology when BLAST E-values were smaller than 10^{-3}. The genome of this novel DV consisted of s 4734 nt with short ITRs of 130 and 77 nucleotides (nt) sequences at the 3' and 5' ends respectively. The Iteravirus genome size is about 5 kb with ITRs of 250 nt suggesting that the ITRs of this novel densovirus are not complete [5]. The genomic organization of this densovirus is monosense with three predicted intron-less ORFs encoding two nonstructural proteins (NS) and one structural protein (VP). ORF1 (nt 253 to 2505) has a coding capacity of 750 amino acids (aa) and contains the typical NS1 helicase superfamily III. ORF2 (nt 2559 to 4568) encodes a 669 aa protein, corresponding to VP and contains the characteristic phospholipase A2 motif [6]. ORF3 (nt 380 to 1729) had a coding capacity for NS2 of 449 aa and typically overlapped NS1. Alignment of the VP and NS protein sequences using Clustal W 1.8.1 [7] revealed that this genome had the highest identity (84,9%) with *Danaus plexippus plexippus densovirus* (DpplDV; GenBank accession number KF963252) [8]. This genome was independently purified from leaves of the original plant stored at -80°C (Qiagen plant DNAeasy kit). PCR products were obtained from different leaves using 15 pairs of primers covering the whole genome that were sequenced using Sanger's method (Cogenics).

58 Recombination analyses using RDP4.18 [9] revealed that this novel DV might
59 result from an intra-genus recombination event between DpplDV and *Dendrolimus*
60 *punctatus* densovirus (DpDV).

61 No insect has been found in any part of this plant and no reads obtained from
62 this plant were assigned to arthropods and no product were obtained using an insect
63 DNA barcoding based on the PCR amplification of a fragment of the mitochondrial
64 cytochrome c oxidase subunit I gene [10]. This densovirus might come from
65 contamination of the plant aerial part by infected arthropods or circulate
66 systemically *in planta* as already reported [11]. This virus was tentatively named
67 *Hordeum marinum densovirus* (HormaDV).

68 **Nucleotide sequence accession number.** The GenBank accession number of
69 HormaDV is KM576800

70 **ACKNOWLEDGMENTS**

71 S. F. was supported by a scholarship from Institut National de la Recherche
72 Agronomique (INRA). P. B. fellowship was funded by the Languedoc-Roussillon
73 Region and the DGA (Département Général des Armées, France). R. F.
74 acknowledges the support of the Centre National de Recherche Scientifique
75 (CNRS) and the Institut de Recherche pour le Développement (IRD).

76 **REFERENCES**

[1] M. Bergoin and P. Tijssen, “Densoviruses : a Highly Diverse Group of Arthropod Parvoviruses,” in *Insect Virology, Asgari, S., Johnson, K.N.*, 2010, pp. 59–72.

[2] S. F. Cotmore, M. Agbandje-McKenna, J. a Chiorini, D. V Mukha, D. J. Pintel, J. Qiu, M. Soderlund-Venermo, P. Tattersall, P. Tijssen, D. Gatherer, and A. J. Davison, “The family Parvoviridae.,” *Arch. Virol.*, vol. 159, no. 5, pp. 1239–47, May 2014.

[3] T. Candresse, D. Filloux, B. Muhire, C. Julian, S. Galzi, G. Fort, P. Bernardo, J.-H. Daugrois, E. Fernandez, D. P. Martin, A. Varsani, and P. Roumagnac, “Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context.,” *PLoS One*, vol. 9, no. 7, p. e102945, Jan. 2014.

[4] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond, “Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.,” *Bioinformatics*, vol. 28, no. 12, pp. 1647–9, Jun. 2012.

[5] Q. Yu and P. Tijssen, “Gene expression of five different iteradensoviruses: BmDV, CeDV, PpDV, SfDV and DpIDV.,” *J. Virol.*, 2014.

[6] Z. Zádori, J. Szelei, M. C. Lacoste, Y. Li, S. Gariépy, P. Raymond, M. Allaire, I. R. Nabi, and P. Tijssen, “A viral phospholipase A2 is required for parvovirus infectivity.,” *Dev. Cell*, vol. 1, no. 2, pp. 291–302, Aug. 2001.

[7] M. a Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. a McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, a Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, “Clustal W and Clustal X version 2.0.,” *Bioinformatics*, vol. 23, no. 21, pp. 2947–8, Nov. 2007.

[8] Q. Yu and P. Tijssen, “Iteradensovirus from the Monarch Butterfly , Danaus plexippus plexippus,” vol. 2, no. 2, pp. 6–7, 2014.

[9] D. P. Martin, P. Lemey, M. Lott, V. Moulton, D. Posada, and P. Lefeuvre, “RDP3: a flexible and fast computer program for analyzing recombination.,” *Bioinformatics*, vol. 26, no. 19, pp. 2462–3, Oct. 2010.

[10] J. J. Wilson, “DNA barcodes for insects.,” in *DNA Barcodes*, Methods Mol Biol., 2012, pp. 17–46

[11] M. van Munster, A. Janssen, A. Clérivet, and J. van den Heuvel, “Can plants use an entomopathogenic virus as a defense against herbivores?,” *Oecologia*, vol. 143, no. 3, pp. 396–401, Apr. 2005.

Références bibliographiques

Acosta-Leal, R., Duffy, S., Xiong, Z., Hammond, R.W., Elena, S.F., 2011. Advances in plant virus evolution: translating evolutionary insights into better disease management. Phytopathology 101(10), 1136-1148.

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Molecular plant pathology 10(4), 537-545.

Adriaenssens, E.M., Van Zyl, L., De Maayer, P., Rubagotti, E., Rybicki, E., Tuffin, M., Cowan, D.A., 2014. Metagenomic analysis of the viral community in Namib Desert hypoliths. Environmental microbiology.

Agindotan, B.O., Ahonsi, M.O., Domier, L.L., Gray, M.E., Bradley, C.A., 2010. Application of sequence-independent amplification (SIA) for the identification of RNA viruses in bioenergy crops. Journal of virological methods 169(1), 119-128.

Agnew, P., J, C.K., Michalakis, Y., 2000. Host life history responses to parasitism. Microbes and infection / Institut Pasteur 2(8), 891-896.

Al Rwahnih, M., Daubert, S., Golino, D., Rowhani, A., 2009. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387(2), 395-401.

Alemandri, V., Rodriguez, P., Izaurralde, J., Medina, S.G., Caro, E.A., Mattio, M.F., Dumon, A., Rodriguez, S.M., Truol, G., 2012. incidence of begomoviruses and climatic characterisation of *Bemisia tabaci*-geminivirus complex in soybean and bean in Argentina. Agriscientia XXIX, 31-39.

Alexander, H., 1998. The interaction between plant competition and disease. Perspectives in Plant Ecology, Evolution and Systematics 1(2), 206-220.

Alexander, H.M., 2010. Disease in Natural Plant Populations, Communities, and Ecosystems: Insights into Ecological and Evolutionary Processes. Plant Dis 94(5), 492-503.

Alexander, H.M., Mauck, K.E., Whitfield, A.E., Garrett, K.A., Malmstrom, C.M., 2013. Plant-virus interactions and the agro-ecological interface. European Journal of Plant Pathology 138(3), 529-547.

Allander, T., Emerson, S.U., Engle, R.E., Purcell, R.H., Bukh, J., 2001. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proceedings of the National Academy of Sciences of the United States of America 98(20), 11609-11614.

Allen, H.K., Bunge, J., Foster, J.A., Bayles, D.O., Stanton, T.B., 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. Microbiome 1(1), 5.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology 215(3), 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25(17), 3389-3402.

Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., Daszak, P., 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in ecology & evolution 19(10), 535-544.

Anderson, R.M., May, R.M., 1982. Coevolution of Hosts and Parasites. Parasitology 85(Oct), 411-426.

Anderson, R.M., May, R.M., 1991. Infectious diseases of humans : dynamics and control, viii, 757 p. pp. Oxford science publications. Oxford University Press, Oxford ; New York.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., Rohwer, F., 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. Bmc Bioinformatics 6, 41.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A., Rohwer, F., 2006. The marine viromes of four oceanic regions. PLoS biology 4(11), e368.

Angly, F.E., Willner, D., Prieto-Davo, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D.A., Barott, K., Cottrell, M.T., Desnues, C., Dinsdale, E.A., Furlan, M., Haynes, M., Henn, M.R., Hu, Y., Kirchman, D.L., McDole, T., McPherson, J.D., Meyer, F., Miller, R.M., Mundt, E., Naviaux, R.K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., Rohwer, F., 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS computational biology 5(12), e1000593.

Anonymous, 1997. Surfer v9.0, Surface Mapping System. Colorado, USA: Golden Software Inc.

Archie, E.A., Luikart, G., Ezenwa, V.O., 2009. Infecting epidemiology with genetics: a new frontier in disease ecology. Trends in ecology & evolution 24(1), 21-30.

Astier, S., 2007. Principles of plant virology : genome, pathogenicity, virus ecology, xxi, 472 p., 478 p. of plates pp. Science Publishers, Enfield, NH.

Baas Becking, L.G.M., 1934. Geobiologie of inleiding tot de milieukunde, The Hague, the Netherlands: W.P. Van Stockum & Zoon

Bailey, T.C., 1994. A review of statistical spatial analysis in geographical information systems. In: Fotheringham, A.S., Rogerson, P.A. (Eds.), Spatial analysis and GIS. Taylor and Francis, London, pp. 13-44.

Bar-Joseph, M., Garnsey, S.M., Gonsalves, D., 1979. The closteroviruses: a distinct group of elongated plant viruses. Adv Virus Res 25, 93-168.

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6(1), 106-136.

Beerenwinkel, N., Gunthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol 3.

Beerenwinkel, N., Zagordi, O., 2011. Ultra-deep sequencing for the analysis of viral populations. Current Opinion in Virology 1(5), 413-418.

Begon, M., Harper, J.L., Townsend, C.R., 1990. Ecology : individuals, populations, and communities, xii, 945 p. pp. 2nd ed. Blackwell Scientific Publications ;Distributors, USA, Publishers' Business Services, Boston Brookline Village, Mass.

Bergelson, J., Dwyer, G., Emerson, J.J., 2001. Models and data on plant-enemy coevolution. Annual review of genetics 35, 469-499.

Bernardo, P., Golden, M., Akram, M., Naimuddin, Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A.G., Peterschmitt, M., Martin, D.P., Roumagnac, P., 2013. Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. Virus research 177(1), 35-45.

Betancourt, M., Fraile, A., Garcia-Arenal, F., 2011. Cucumber mosaic virus satellite RNAs that induce similar symptoms in melon plants show large differences in fitness. Journal of General Virology 92, 1930-1938.

Bibby, K., 2013. Metagenomic identification of viral pathogens. Trends in biotechnology 31(5), 275-279.

Biek, R., Real, L.A., 2010. The landscape genetics of infectious disease emergence and spread. Mol Ecol 19(17), 3515-3531.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., Taylor, J., 2010. Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [*et al.*] Chapter 19, Unit 19 10 11-21.

Blinkova, O., Kapoor, A., Victoria, J., Jones, M., Wolfe, N., Naeem, A., Shaukat, S., Sharif, S., Alam, M.M., Angez, M., Zaidi, S., Delwart, E.L., 2009. Cardioviruses are genetically diverse and cause common enteric infections in South Asian children. Journal of virology 83(9), 4631-4641.

Blinkova, O., Victoria, J., Li, Y., Keele, B.F., Sanz, C., Ndjango, J.B., Peeters, M., Travis, D., Lonsdorf, E.V., Wilson, M.L., Pusey, A.E., Hahn, B.H., Delwart, E.L., 2010. Novel circular DNA viruses in stool samples of wild-living chimpanzees. The Journal of general virology 91(Pt 1), 74-86.

Blomstrom, A.L., 2011. Viral metagenomics as an emerging and powerful tool in veterinary medicine. The Veterinary quarterly 31(3), 107-114.

Borer, E.T., Hosseini, P.R., Seabloom, E.W., Dobson, A.P., 2007. Pathogen-induced reversal of native dominance in a grassland community. Proceedings of the National Academy of Sciences of the United States of America 104(13), 5473-5478.

Borer, E.T., Mitchell, C.E., Power, A.G., Seabloom, E.W., 2009. Consumers indirectly increase infection risk in grassland food webs. Proceedings of the National Academy of Sciences of the United States of America 106(2), 503-506.

Bragard, C., Caciagli, P., Lemaire, O., Lopez-Moya, J.J., MacFarlane, S., Peters, D., Susi, P., Torrance, L., 2013. Status and prospects of plant virus control through interference with vector transmission. Annual review of phytopathology 51, 177-201.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B., Mahaffy, J.M., Mueller, J., Nulton, J., Rayhawk, S., Rodriguez-Brito, B., Salamon, P., Rohwer, F., 2008. Viral diversity and dynamics in an infant gut. Research in microbiology 159(5), 367-373.

Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends in microbiology 13(6), 278-284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer, F., 2002. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99(22), 14250-14255.

Briddon, R.W., Bedford, I.D., Tsai, J.H., Markham, P.G., 1996. Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology 219(2), 387-394.

Brown, J.K., Tellier, A., 2011. Plant-parasite coevolution: bridging the gap between genetics and ecology. Annual review of phytopathology 49, 345-367.

Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., Coutinho, P.M., Henrissat, B., Nelson, K.E., White, B.A., 2009. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proceedings of the National Academy of Sciences of the United States of America 106(6), 1948-1953.

Bull, R.A., Eden, J.S., Luciani, F., McElroy, K., Rawlinson, W.D., White, P.A., 2012. Contribution of Intra- and Interhost Dynamics to Norovirus Evolution. Journal of virology 86(6), 3219-3229.

Bunge, J., 2011. Estimating the number of species with CatchAll. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 121-130.

Bunge, J., Willis, A., Walsh, F., 2014. Estimating the Number of Species in Microbial Diversity Studies. Annu Rev Stat Appl 1, 427-445.

Bunge, J., Woodard, L., Bohning, D., Foster, J.A., Connolly, S., Allen, H.K., 2012. Estimating population diversity with CatchAll. Bioinformatics 28(7), 1045-1047.

Burdon, J.J., 1987. Diseases and plant population biology, viii, 208 p. pp. Cambridge studies in ecology. Cambridge University Press, Cambridge Cambridgeshire ; New York.

Burdon, J.J., Chilvers, G.A., 1982. Host Density as a Factor in Plant-Disease Ecology. Annual review of phytopathology 20, 143-166.

Burdon, J.J., Thrall, P.H., 2008. Pathogen evolution across the agro-ecological interface: implications for disease management. Evolutionary Applications 1(1), 57-65.

Burdon, J.J., Thrall, P.H., 2013. What have we learned from studies of wild plant-pathogen associations?—the dynamic interplay of time, space and life-history. European Journal of Plant Pathology 138(3), 417-429.

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.H., Fernandez, E., Martin, D.P., Varsani, A., Roumagnac, P., 2014. Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PloS one 9(7), e102945.

Cann, A.J., 2012. Principles of Molecular Virology, 5th Edition. Principles of Molecular Virology, 5th Edition, 1-303.

Canuti, M., van der Hoek, L., 2014. Virus discovery: are we scientists or genome collectors? Trends in microbiology 22(5), 229-231.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7(5), 335-336.

Caranta, C., Lefebvre, V., Palloix, A., 1997. Polygenic resistance of pepper to potyviruses consists of a combination of isolate-specific and broad-spectrum quantitative trait loci. Mol Plant Microbe In 10(7), 872-878.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.-J., 2004. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett 8(2), 148-159.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.J., 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62(2), 361-371.

Chao, A., Shen, T.J., 2010. Program SPADE (Species Prediction And Diversity Estimation).

Charuvaka, A., Rangwala, H., 2011. Evaluation of short read metagenomic assembly. Bmc Genomics 12 Suppl 2, S8.

Chave, J., 2013. The problem of pattern and scale in ecology: what have we learned in 20years? Ecol Lett 16, 4-16.

Clarke, D.D., 1986. Tolerance of parasites and disease in plants and its significance in host-parasite interactions. Adv. Plant. Pathol. 5, 161-198.

Clover, G.R.G., Smith, H.G., Azam-Ali, S.N., Jaggard, K.W., 1999. The effects of drought on sugar beet growth in isolation and in combination with beet yellows virus infection. J Agr Sci 133, 251-261.

Coakley, S.M., Scherm, H., Chakraborty, S., 1999. Climate change and plant disease management. Annual review of phytopathology 37, 399-426.

Coetzee, B., Freeborough, M.J., Maree, H.J., Celton, J.M., Rees, D.J.G., Burger, J.T., 2010. Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology 400(2), 157-163.

Cooper, I., Jones, R.A.C., 2006. Wild plants and viruses: Under-investigated ecosystems. Adv Virus Res 67, 1-47.

Corbin, J.D., D'Antonio, C.M., 2004. Competition between native perennial and exotic annual grasses: Implications for an historical invasion. Ecology 85(5), 1273-1283.

Costa, A.S., 1975. Increase in the population density of Bemisia tabaci, a threat of widespread virus infection of legume crops in Brazil. In: Bird, J., Maramorosch, K. (Eds.), Tropical Diseases of Legumes. Academic Press, pp. 27–49.

Coutts, B.A., Thomas-Carroll, M.L., Jones, R.A.C., 2004. Analysing spatial patterns of spread of Lettuce necrotic yellows virus and lettuce big-vein disease in lettuce field plantings. Ann Appl Biol 145(3), 339-343.

Creager, A.N., Scholthof, K.B., Citovsky, V., Scholthof, H.B., 1999. Tobacco mosaic virus. Pioneering research for a century. The Plant cell 11(3), 301-308.

Culley, A.I., Lang, A.S., Suttle, C.A., 2007. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. Virology journal 4.

Dader, B., Moreno, A., Vinuela, E., Fereres, A., 2012. Spatio-temporal dynamics of viruses are differentially affected by parasitoids depending on the mode of transmission. Viruses 4(11), 3069-3089.

Daszak, P., Cunningham, A.A., Hyatt, A.D., 2000. Emerging infectious diseases of wildlife--threats to biodiversity and human health. Science 287(5452), 443-449.

Dayaram, A., Opong, A., Jaschke, A., Hadfield, J., Baschiera, M., Dobson, R.C.J., Offei, S.K., Shepherd, D.N., Martin, D.P., Varsani, A., 2012. Molecular characterisation of a novel cassava associated circular ssDNA virus. Virus research 166(1-2), 130-135.

De Barro, P.J., Liu, S.S., Boykin, L.M., Dinsdale, A.B., 2011. Bemisia tabaci: a statement of species status. Annual review of entomology 56, 1-19.

De Wit, C.T., 1986. On competition, Evolutionary Monographs, University of Chicago.

Decker, C.J., Parker, R., 2014. Analysis of Double-Stranded RNA from Microbial Communities Identifies Double-Stranded RNA Virus-like Elements. Cell Rep 7(3), 898-906.

Degnan, P.H., Ochman, H., 2012. Illumina-based analysis of microbial community diversity. The ISME journal 6(1), 183-194.

Delatte, H., Reynaud, B., Granier, M., Thornary, L., Lett, J.M., Goldbach, R., Peterschmitt, M., 2007. A new silverleaf-inducing biotype Ms of Bemisia tabaci (Hemiptera: Aleyrodidae) indigenous to the islands of the south-west Indian Ocean. Bulletin of Entomological Research 95(01).

Dellaporta, S., L., Wood, J., Hicks, J., B., 1983. A plant DNA minipreparation: Version II. Plant Molecular Biology Reporter 1(4), 19-21.

Delwart, E.L., 2007. Viral metagenomics. Rev Med Virol 17(2), 115-131.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M.A., Nelson, K.E., Nilsson, C., Olson, R., Paul, J., Brito, B.R., Ruan, Y., Swan, B.K., Stevens, R., Valentine, D.L., Thurber, R.V., Wegley, L., White, B.A., Rohwer, F., 2008. Functional metagenomic profiling of nine biomes. Nature 452(7187), 629-632.

Dixon, A.F.G., 1985. Aphid ecology, ix, 157 p. pp. Blackie ;Distributed in the U.S.A. by Chapman and Hall, Glasgow New York.

Dixon, P., 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14(6), 927-930.

Dodds, J.A., Morris, T.J., Jordan, R.L., 1984. Plant Viral Double-Stranded-Rna. Annual review of phytopathology 22, 151-168.

Domingo-Calap, P., Cuevas, J.M., Sanjuan, R., 2009. The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. PLoS genetics 5(11), e1000742.

Donaire, L., Wang, Y., Gonzalez-Ibeas, D., Mayer, K.F., Aranda, M.A., Llave, C., 2009. Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392(2), 203-214.

Donaldson, E.F., Haskew, A.N., Gates, J.E., Huynh, J., Moore, C.J., Frieman, M.B., 2010. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. Journal of virology 84(24), 13004-13018.

Dong, H., Chen, Y.Y., Shen, Y., Wang, S.Y., Zhao, G.P., Jin, W.R., 2011. Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System. Acta Bioch Bioph Sin 43(6), 496-500.

Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proceedings of the National Academy of Sciences of the United States of America 88(16), 7160-7164.

Droge, J., McHardy, A.C., 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Briefings in bioinformatics 13(6), 646-655.

Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A., He, Z., 2014. Identification and molecular characterization of a single-stranded circular DNA virus with similarities to Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1. Archives of virology 159(6), 1527-1531.

Duffy, S., Turner, P.E., Burch, C.L., 2006. Pleiotropic costs of niche expansion in the RNA bacteriophage phi 6. Genetics 172(2), 751-757.

Dunn, A.M., Torchin, M.E., Hatcher, M.J., Kotanen, P.M., Blumenthal, D.M., Byers, J.E., Coon, C.A.C., Frankel, V.M., Holt, R.D., Hufbauer, R.A., Kanarek, A.R., Schierenbeck, K.A., Wolfe, L.M., Perkins, S.E., 2012. Indirect effects of parasites in invasions. Funct Ecol 26(6), 1262-1274.

Dunne, W.M., Jr., Westblade, L.F., Ford, B., 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology.

Dybdahl, M.F., Storfer, A., 2003. Parasite local adaptation: Red Queen versus Suicide King. Trends in ecology & evolution 18(10), 523-530.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19), 2460-2461.

Edwards, R., 1996. Tomorrow's bitter harvest. New Sci 151(2043), 14-15.

Elena, S.F., Fraile, A., Garcia-Arenal, F., 2014. Evolution and emergence of plant viruses. Adv Virus Res 88, 161-191.

Ellis, E.C., Klein Goldewijk, K., Siebert, S., Lightman, D., Ramankutty, N., 2010. Anthropogenic transformation of the biomes, 1700 to 2000. Global Ecology and Biogeography, no-no.

Ellis, E.C., Ramankutty, N., 2008. Putting people in the map: anthropogenic biomes of the world. Frontiers in Ecology and the Environment 6(8), 439-447.

Ellstrand, N.C., 2003. Current knowledge of gene flow in plants: implications for transgene flow. Philos T R Soc B 358(1434), 1163-1170.

Ewald, P.W., 1983. Host-Parasite Relations, Vectors, and the Evolution of Disease Severity. Annu Rev Ecol Syst 14, 465-485.

Fabre, F., Montarry, J., Coville, J., Senoussi, R., Simon, V., Moury, B., 2012. Modelling the evolutionary dynamics of viruses within their hosts: a case study using high-throughput sequencing. PLoS pathogens 8(4), e1002654.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M., Thresh, J.M., 2006. Molecular ecology and emergence of tropical plant viruses. Annual review of phytopathology 44, 235-260.

Fauquet, C.M., Briddon, R.W., Brown, J.K., Moriones, E., Stanley, J., Zerbini, M., Zhou, X., 2008. Geminivirus strain demarcation and nomenclature. Archives of virology 153(4), 783-821.

Fereres, A., Moreno, A., 2009. Behavioural aspects influencing plant virus transmission by homopteran insects. Virus research 141(2), 158-168.

Ferris, M.T., Joyce, P., Burch, C.L., 2007. High frequency of mutations that expand the host range of an RNA virus. Genetics 176(2), 1013-1022.

Flory, S.L., Clay, K., 2013. Pathogen accumulation and long-term dynamics of plant invasions. J Ecol 101(3), 607-613.

Forterre, P., 2010. Defining life: the virus viewpoint. Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life 40(2), 151-160.

Fraile, A., Garcia-Arenal, F., 2010. The coevolution of plants and viruses: resistance and pathogenicity. Adv Virus Res 76, 1-32.

Frank, S.A., 1996. Models of parasite virulence. Q Rev Biol 71(1), 37-78.

Froussard, P., 1993. rPCR: a powerful tool for random amplification of whole RNA sequences. PCR methods and applications 2(3), 185-190.

Gallet, R., Bonnot, F., Milazzo, J., Tertois, C., Adreit, H., Ravigne, V., Tharreau, D., Fournier, E., 2013. The variety mixture strategy assessed in a G x G experiment with rice and the blast fungus Magnaporthe oryzae. Frontiers in genetics 4, 312.

Garcia-Andres, S., Monci, F., Navas-Castillo, J., Moriones, E., 2006. Begomovirus genetic diversity in the native plant reservoir Solanum nigrum: Evidence for the presence of a new virus species of recombinant nature. Virology 350(2), 433-442.

Garcia-Arenal, F., Fraile, A., Malpica, J.M., 2001. Variability and genetic structure of plant virus populations. Annual review of phytopathology 39, 157-186.

Garrett, K.A., Dendy, S.P., Frank, E.E., Rouse, M.N., Travers, S.E., 2006. Climate change effects on plant disease: genomes to ecosystems. Annual review of phytopathology 44, 489-509.

Garrett, K.A., Nita, M., DeWolf, E. D., Gomez, L., and Sparks, A. H. , 2009. Plant pathogens as indicators of climate change. Climate change: Observed Impacts on Planet Earth, Letcher, T., Elseiver, Dordedrecht, 425-437.

Germain, J.F., Chatot, C., Meusnier, I., Artige, E., Rasplus, J.Y., Cruaud, A., 2013. Molecular identification of Epitrix potato flea beetles (Coleoptera: Chrysomelidae) in Europe and North America. Bull Entomol Res 103(3), 354-362.

Ghosh, T.S., Mohammed, M.H., Komanduri, D., Mande, S.S., 2011. ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. Bioinformation 6(2), 91-94.

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., Terlizzi, F., Credi, R., Saldarelli, P., 2012. A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus research 163(1), 262-268.

Gibbs, A., 1980. A plant virus that partially protects its wild legume host against herbivores. Intervirology 13(1), 42-47.

Gibbs, A.J., Gibbs, M.J., 2006. A broader definition of 'the virus species'. Archives of virology 151(7), 1419-1422.

Gibbs, M.J., Smeianov, V.V., Steele, J.L., Upcroft, P., Efimov, B.A., 2006. Two families of Rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol Biol Evol 23(6), 1097-1100.

Gilbert, G.S., 2002. Evolutionary ecology of plant diseases in natural ecosystems. Annual review of phytopathology 40, 13-43.

Goecks, J., Nekrutenko, A., Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 11(8), R86.

Goldbach, R., Peters, D., 1994. Possible Causes of the Emergence of Tospovirus Diseases. Semin Virol 5(2), 113-120.

Goll, J., Rusch, D.B., Tanenbaum, D.M., Thiagarajan, M., Li, K., Methe, B.A., Yooseph, S., 2010. METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics. Bioinformatics 26(20), 2631-2632.

Gomez-Alvarez, V., Teal, T.K., Schmidt, T.M., 2009. Systematic artifacts in metagenomes from complex microbial communities. The ISME journal 3(11), 1314-1317.

Greger, M., 2007. The human/animal interface: emergence and resurgence of zoonotic infectious diseases. Critical reviews in microbiology 33(4), 243-299.

Grigoras, I., Ginzo, A.I., Martin, D.P., Varsani, A., Romero, J., Mammadov, A., Huseynova, I.M., Aliyev, J.A., Kheyr-Pour, A., Huss, H., Ziebell, H., Timchenko, T., Vetten, H.J., Gronenborn, B., 2014. Genome diversity and evidence of recombination and reassortment in nanoviruses from Europe. The Journal of general virology 95(Pt 5), 1178-1191.

Guindon, S., Lethiec, F., Duroux, P., Gascuel, O., 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic acids research 33(Web Server issue), W557-559.

Hafez, E.E., Aseel, D.G., Mostafa, S., 2014. Two Novel Mycoviruses Related to Geminivirus Isolated from the Soil-Borne FungiMacrophomina Phaseolina(Tassi) Goid. andMucor RacemosusBull. Biotechnology & Biotechnological Equipment 27(6), 4222-4226.

Hall, R.J., Leblanc-Maridor, M., Wang, J., Ren, X., Moore, N.E., Brooks, C.R., Peacey, M., Douwes, J., McLean, D.J., 2013. Metagenomic detection of viruses in aerosol samples from workers in animal slaughterhouses. PloS one 8(8), e72226.

Hall, R.J., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S., Strydom, H., Moore, N.E., Ren, X., Huang, Q.S., Carter, P.E., Peacey, M., 2014. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. Journal of virological methods 195, 194-204.

Hamady, M., Knight, R., 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 19(7), 1141-1152.

Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2001. PAST: Paleontological statistics software package for education and data analysis, Palaeontologia Electronica. Vol. 4, pp. 9pp.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chem Biol 5(10), R245-R249.

Hanski, I., Gilpin, M., 1991. Metapopulation Dynamics - Brief-History and Conceptual Domain. Biol J Linn Soc 42(1-2), 3-16.

Harrison, B., Robinson, D., 1999. Natural Genomic and Antigenic Variation in Whitefly-Transmitted Geminiviruses (Begomoviruses). Annual review of phytopathology 37, 369-398.

Harvell, C.D., Mitchell, C.E., Ward, J.R., Altizer, S., Dobson, A.P., Ostfeld, R.S., Samuel, M.D., 2002. Climate warming and disease risks for terrestrial and marine biota. Science 296(5576), 2158-2162.

Herms, D.A., Mattson, W.J., 1992. The Dilemma of Plants - to Grow or Defend. Q Rev Biol 67(3), 283-335.

Hogenhout, S.A., Ammar el, D., Whitfield, A.E., Redinbaugh, M.G., 2008. Insect vector interactions with persistently transmitted viruses. Annual review of phytopathology 46, 327-359.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacon, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y.D., Lahaye, R., Lee, H.L., Long, D.G., Madrinan, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.K., Little, D.P., Grp, C.P.W., 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America 106(31), 12794-12797.

Holt, R.D., Grover, J., Tilman, D., 1994. Simple Rules for Interspecific Dominance in Systems with Exploitative and Apparent Competition. American Naturalist 144(5), 741-771.

Holt, R.D., Pickering, J., 1985. Infectious-Disease and Species Coexistence - a Model of Lotka-Volterra Form. American Naturalist 126(2), 196-211.

Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography, xiv, 375 p. pp. Monographs in population biology. Princeton University Press, Princeton.

Hull, R., 2002. Matthews' plant virology, xx, 1001 p. pp. 4th ed. Academic Press, San Diego.

Hull, R., 2009. Comparative plant virology, xvi, 376 p. pp. 2nd ed. Elsevier/Academic Press, Amsterdam ; Boston.

Hunter, C.I., Mitchell, A., Jones, P., McAnulla, C., Pesseat, S., Scheremetjew, M., Hunter, S., 2012. Metagenomic analysis: the challenge of the data bonanza. Briefings in bioinformatics 13(6), 743-746.

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res 17(3), 377-386.

Janzen, D.H., 1980. When Is It Coevolution. Evolution 34(3), 611-612.

Jeske, H., 2009. Geminiviruses. Curr Top Microbiol 331, 185-226.

Johne, R., Muller, H., Rector, A., van Ranst, M., Stevens, H., 2009. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends in microbiology 17(5), 205-211.

Jones, R.A., 2013. Trends in plant virus epidemiology: Opportunities from new or improved technologies. Virus research.

Jones, R.A.C., 2005. Patterns of spread of two non-persistently aphid-borne viruses in lupin stands under four different infection scenarios. Ann Appl Biol 146(3), 337-350.

Jones, R.A.C., 2009. Plant virus emergence and evolution: Origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus research 141(2), 113-130.

Jones, R.A.C., Coutts, B.A., Latham, L.J., McKirdy, S.J., 2008. Cucumber mosaic virus infection of chickpea stands: temporal and spatial patterns of spread and yield-limiting potential. Plant Pathol 57(5), 842-853.

Jousimo, J., Tack, A.J., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., Laine, A.L., 2014. Disease ecology. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. Science 344(6189), 1289-1293.

Kareiva, P., Watts, S., McDonald, R., Boucher, T., 2007. Domesticated nature: shaping landscapes and ecosystems for human welfare. Science 316(5833), 1866-1869.

Kashif, M., Pietila, S., Artola, K., Jones, R.A.C., Tugume, A.K., Makinen, V., Valkonen, J.P.T., 2012. Detection of Viruses in Sweetpotato from Honduras and Guatemala Augmented by Deep-Sequencing of Small-RNAs. Plant Dis 96(10), 1430-1437.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12), 1647-1649.

Keesing, F., Belden, L.K., Daszak, P., Dobson, A., Harvell, C.D., Holt, R.D., Hudson, P., Jolles, A., Jones, K.E., Mitchell, C.E., Myers, S.S., Bogich, T., Ostfeld, R.S., 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature 468(7324), 647-652.

Keesing, F., Holt, R.D., Ostfeld, R.S., 2006. Effects of species diversity on disease risk. Ecol Lett 9(4), 485-498.

Kehoe, M.A., Coutts, B.A., Buirchell, B.J., Jones, R.A., 2014. Hardenbergia mosaic virus: crossing the barrier between native and introduced plant species. Virus research 184, 87-92.

Kelly, D.W., Paterson, R.A., Townsend, C.R., Poulin, R., Tompkins, D.M., 2009. Parasite spillback: a neglected concept in invasion ecology? Ecology 90(8), 2047-2056.

Kennedy, G.G., Barbour, J.D., 1992. Resistance variation in natural and managed systems. In: Fritz, R.S., Sims, E.L. (Ed.), Plant resistance to herbivores and pathogens: Ecology, evolution, and genetics. University of Chicago Press, Chicago, IL, pp. 13-41.

Khetarpal, R.K., Maury, Y., 1987. Pea seed-borne mosaic virus: A review. Agronomie 7, 215-224.

Kim, K.H., Chang, H.W., Nam, Y.D., Roh, S.W., Kim, M.S., Sung, Y., Jeon, C.O., Oh, H.M., Bae, J.W., 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Applied and environmental microbiology 74(19), 5975-5985.

Kitron, U., 1998. Landscape ecology and epidemiology of vector-borne diseases: tools for spatial analysis. Journal of medical entomology 35(4), 435-445.

Kraaij, T., 2010. Changing the fire management regime in the renosterveld and lowland fynbos of the Bontebok National Park. S Afr J Bot 76(3), 550-557.

Krawetz, S.A., 1989. Sequence errors described in GenBank: a means to determine the accuracy of DNA sequence interpretation. Nucleic acids research 17(10), 3951-3957.

Kress, W.J., Erickson, D.L., Jones, F.A., Swenson, N.G., Perez, R., Sanjur, O., Bermingham, E., 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences of the United States of America 106(44), 18621-18626.

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. Virology 388(1), 1-7.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., Hugenholtz, P., 2008. A Bioinformatician's Guide to Metagenomics. Microbiol Mol Biol R 72(4), 557-578.

Labonte, J.M., Suttle, C.A., 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. The ISME journal 7(11), 2169-2177.

Laserson, J., Jojic, V., Koller, D., 2011. Genovo: de novo assembly for metagenomes. Journal of computational biology : a journal of computational molecular cell biology 18(3), 429-443.

Lecoq, H., Ravelonandro, M., Wipf-Scheibel, C., Monsion, M., Raccah, B., Dunez, J., 1994. Significance of the heterologous encapsidation of zucchini yellow mosaic potyvirus in transgenic plants expressing plum pox potyvirus capsid protein. EPPO Bulletin 24, 555-559.

Lefeuvre, P., Lett, J.M., Reynaud, B., Martin, D.P., 2007a. Avoidance of protein fold disruption in natural virus recombinants. PLoS pathogens 3(11), e181.

Lefeuvre, P., Martin, D.P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B., Lett, J.M., 2007b. Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. The Journal of general virology 88(Pt 12), 3458-3468.

Levin, R.A., Wagner, W.L., Hoch, P.C., Nepokroeff, M., Pires, J.C., Zimmer, E.A., Sytsma, K.J., 2003. Family-level relationships of Onagraceae based on chloroplast rbcL and ndhF data. American journal of botany 90(1), 107-115.

Levins, R., 1968. Evolution in changing environments; some theoretical explorations, ix, 120 p. pp. Monographs in population biology,. Princeton University Press, Princeton, N.J.,.

Lin, K.Y., Cheng, C.P., Chang, B.C., Wang, W.C., Huang, Y.W., Lee, Y.S., Huang, H.D., Hsu, Y.H., Lin, N.S., 2010. Global analyses of small interfering RNAs derived from Bamboo mosaic virus and its associated satellite RNAs in different plants. PloS one 5(8), e11928.

Lipkin, W.I., 2010. Microbe hunting. Microbiology and molecular biology reviews : MMBR 74(3), 363-377.

Little, T.J., Shuker, D.M., Colegrave, N., Day, T., Graham, A.L., 2010. The coevolution of virulence: tolerance in perspective. PLoS pathogens 6(9), e1001006.

Liu, H., Fu, Y., Xie, J., Cheng, J., Ghabrial, S.A., Li, G., Peng, Y., Yi, X., Jiang, D., 2012a. Evolutionary genomics of mycovirus-related dsRNA viruses reveals cross-family horizontal gene transfer and evolution of diverse viral lineages. Bmc Evol Biol 12, 91.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012b. Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology 2012, 251364.

Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G.P., Saponari, M., 2012. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. Virology 432(1), 162-172.

Lovisolo, O., Hull, R., Rosler, O., 2003. Coevolution of viruses with hosts and vectors and possible paleontology. Advances in Virus Research, Vol 62 62, 325-379.

Lowry, E., 2007. The role of aphid host preference in barley yellow dwarf virus epidemiology, Cornell University.

Lozupone, C., Hamady, M., Knight, R., 2006. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. Bmc Bioinformatics 7, 371.

MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., Roossinck, M., 2013. Biosecurity implications of new technology and discovery in plant virus research. PLoS pathogens 9(8), e1003337.

Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M., Bazzaz, F.A., 2000. Biotic invasions: Causes, epidemiology, global consequences, and control. Ecol Appl 10(3), 689-710.

Madden, L.V., Jeger, M.J., van den Bosch, F., 2000. A theoretical assessment of the effects of vector-virus transmission mechanism on plant virus disease epidemics. Phytopathology 90(6), 576-594.

Malmstrom, C., Stoner, C., Brandenburg, S., Newton, L., 2006. Virus infection and grazing exert counteracting influences on survivorship of native bunchgrass seedlings competing with invasive exotics. J Ecol 94(2), 264-275.

Malmstrom, C.M., Hughes, C.C., Newton, L.A., Stoner, C.J., 2005a. Virus infection in remnant native bunchgrasses from invaded California grasslands. The New phytologist 168(1), 217-230.

Malmstrom, C.M., McCullough, A.J., Johnson, H.A., Newton, L.A., Borer, E.T., 2005b. Invasive annual grasses indirectly increase virus incidence in California native perennial bunchgrasses. Oecologia 145(1), 153-164.

Malmstrom, C.M., Melcher, U., Bosque-Perez, N.A., 2011. The expanding field of plant virus ecology: historical foundations, knowledge gaps, and research directions. Virus research 159(2), 84-94.

Malpica, J.M., Sacristan, S., Fraile, A., Garcia-Arenal, F., 2006. Association and host selectivity in multi-host pathogens. PloS one 1, e41.

Mangla, S., Inderjit, Callaway, R.M., 2008. Exotic invasive plant accumulates native soil pathogens which inhibit native plants. J Ecol 96(1), 58-67.

Mansky, L.M., Cunningham, K.S., 2000. Virus mutators and antimutators: roles in evolution, pathogenesis and emergence. Trends in genetics : TIG 16(11), 512-517.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer research 27(2), 209-220.

Mardia, K.V., Goodall, C.R., 1993. Spatial-Temporal Analysis of Multivariate Environmental Monitoring Data. N-Holland Stat Prob 6, 347-386.

Marquez, L.M., Redman, R.S., Rodriguez, R.J., Roossinck, M.J., 2007. A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. Science 315(5811), 513-515.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011a. Recombination in eukaryotic single stranded DNA viruses. Viruses 3(9), 1699-1738.

Martin, D.P., Lefeuvre, P., Varsani, A., Hoareau, M., Semegni, J.Y., Dijoux, B., Vincent, C., Reynaud, B., Lett, J.M., 2011b. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS pathogens 7(9), e1002203.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26(19), 2462-2463.

Martin, D.P., Willment, J.A., Rybicki, E.P., 1999. Evaluation of Maize Streak Virus Pathogenicity in Differentially Resistant Zea mays Genotypes. Phytopathology 89(8), 695-700.

Martin, R.R., Zhou, J., Tzanetakis, I.E., 2011c. Blueberry latent virus: an amalgam of the Partitiviridae and Totiviridae. Virus research 155(1), 175-180.

Martin, S., Elena, S.F., 2009. Application of game theory to the interaction between plant viruses during mixed infections. Journal of General Virology 90, 2815-2820.

Martiniere, A., Bak, A., Macia, J.L., Lautredou, N., Gargani, D., Doumayrou, J., Garzo, E., Moreno, A., Fereres, A., Blanc, S., Drucker, M., 2013. A virus responds instantly to the presence of the vector on the host and forms transmission morphs. eLife 2, e00183.

Maule, A.J., Caranta, C., Boulton, M.I., 2007. Sources of natural resistance to plant viruses: status and prospects. Molecular plant pathology 8(2), 223-231.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., Kyrpides, N.C., 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature methods 4(6), 495-500.

McCallum, H., Barlow, N., Hone, J., 2001. How should pathogen transmission be modelled? Trends in ecology & evolution 16(6), 295-300.

Mckirdy, S.J., Coutts, B.A., Jones, R.A.C., 1994. Occurrence of Bean Yellow Mosaic-Virus in Subterranean Clover Pastures and Perennial Native Legumes. Aust J Agr Res 45(1), 183-194.

Meentemeyer, R.K., Haas, S.E., Vaclavik, T., 2012. Landscape epidemiology of emerging infectious diseases in natural and human-altered ecosystems. Annual review of phytopathology 50, 379-402.

Melcher, U., Muthukumar, V., Wiley, G.B., Min, B.E., Palmer, M.W., Verchot-Lubicz, J., Ali, A., Nelson, R.S., Roe, B.A., Thapa, V., Pierce, M.L., 2008. Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. Journal of virological methods 152(1-2), 49-55.

Melcher, U., Verma, R., Schneider, W.L., 2014. Metagenomic search strategies for interactions among plants and multiple microbes. Frontiers in plant science 5, 268.

Mende, D.R., Waller, A.S., Sunagawa, S., Jarvelin, A.I., Chan, M.M., Arumugam, M., Raes, J., Bork, P., 2012. Assessment of metagenomic assembly using simulated next generation sequencing data. PloS one 7(2), e31386.

Mesleard, F., Mauchamp, A., Pineau, O., Dutoit, T., 2011. Rabbits are more effective than cattle for limiting shrub colonization in Mediterranean xero-halophytic meadows. Ecoscience 18(1), 37-41.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A., 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. Bmc Bioinformatics 9.

Milton, S.J., 2004. Grasses as invasive alien plants in South Africa. South African Journal of Science 100(1), 69-75.

Min, B.E., Feldman, T.S., Ali, A., Wiley, G., Muthukumar, V., Roe, B.A., Roossinck, M., Melcher, U., Palmer, M.W., Nelson, R.S., 2012. Molecular characterization, ecology, and epidemiology of a novel Tymovirus in Asclepias viridis from Oklahoma. Phytopathology 102(2), 166-176.

Mink, G.I., 1993. Pollen and seed-transmitted viruses and viroids. Annual review of phytopathology 31, 375-402.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., Bushman, F.D., 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21(10), 1616-1625.

Mirik, M., Jones, D.C., Price, J.A., Workneh, F., Ansley, R.J., Rush, C.M., 2011. Satellite Remote Sensing of Wheat Infected by Wheat streak mosaic virus. Plant Dis 95(1), 4-12.

Mitchell, C.E., Power, A.G., 2003. Release of invasive plants from fungal and viral pathogens. Nature 421(6923), 625-627.

Mitchell, C.E., Power, A.G., 2006. Plant communities and disease ecology. In: Collinge, S.K., Ray, C. (Eds.), Disease Ecology: Community Structure and Pathogen Dynamics. University Press, Oxford, Oxford, pp. 58–72.

Moffat, A.S., 1999. Plant pathology - Geminiviruses emerge as serious crop threat. Science 286(5446), 1835-1835.

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2(1), 63-77.

Molina, L.G., da Fonseca, G.C., de Morais, G.L., de Oliveira, L.F.V., de Carvalho, J.B., Kulcheski, F.R., Margis, R., 2012. Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries. Genet Mol Biol 35(1), 292-U208.

Monjane, A.L., Harkins, G.W., Martin, D.P., Lemey, P., Lefeuvre, P., Shepherd, D.N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E.K., Lakoutene, D.P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J.B., van Antwerpen, T., Owor, B.E., Flett, B., Ramusi, M., Windram, O.P., Syed, R., Lett, J.M., Briddon, R.W., Markham, P.G., Rybicki, E.P., Varsani, A., 2011. Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. Journal of virology 85(18), 9623-9636.

Moore, S.M., Borer, E.T., 2012. The influence of host diversity and composition on epidemiological patterns at multiple spatial scales. Ecology 93(5), 1095-1105.

Morales, F.J., Anderson, P.K., 2001. The emergence and dissemination of whitefly-transmitted geminiviruses in Latin America - Brief review. Archives of virology 146(3), 415-441.

Moreira, D., Lopez-Garcia, P., 2009. Ten reasons to exclude viruses from the tree of life. Nature Reviews Microbiology 7(4), 306-311.

Moreno, A., Nebreda, M., Diaz, B.M., Garcia, M., Salas, F., Fereres, A., 2007. Temporal and spatial spread of Lettuce mosaic virus in lettuce crops in central Spain: factors involved in Lettuce mosaic virus epidemics. Ann Appl Biol 150(3), 351-360.

Moury, B., Fabre, F., Montarry, J., Janzac, B, Ayme, V., Palloix, A., 2010. L'adaptation des virus de plantes aux résistances variétales. Virologie 14(4), 227-239.

Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V.G., Briddon, R.W., Varsani, A., 2013. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Archives of virology.

Mundt, C.C., 2002. Use of multiline cultivars and cultivar mixtures for disease management. Annual review of phytopathology 40, 381-410.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G.B., Roe, B.A., Palmer, M.W., Thapa, V., Ali, A., Ding, T., 2009. Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. Virus research 141(2), 169-173.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403(6772), 853-858.

Nadler, S.A., 1995. Microevolution and the genetic structure of parasite populations. The Journal of parasitology 81(3), 395-403.

Namiki, T., Hachiya, T., Tanaka, H., Sakakibara, Y., 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic acids research 40(20), e155.

Nancarrow, N., Constable, F.E., Finlay, K.J., Freeman, A.J., Rodoni, B.C., Trebicki, P., Vassiliadis, S., Yen, A.L., Luck, J.E., 2014. The effect of elevated temperature on Barley yellow dwarf virus-PAV in wheat. Virus research 186, 97-103.

Navarro, B., Pantaleo, V., Gisel, A., Moxon, S., Dalmay, T., Bisztray, G., Di Serio, F., Burgyan, J., 2009. Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. PloS one 4(11), e7686.

Nelson, M.R., Orum, T.V., Jaime-Garcia, R., Nadeem, A., 1999. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. Plant Dis 83(4), 308-319.

Ng, T.F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B.S., Wommack, K.E., Delwart, E., 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. Journal of virology.

Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011a. Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. PloS one 6(4).

Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y.J., Rohwer, F., Breitbart, M., 2011b. Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes. PloS one 6(6).

Olson, A.J., Pataky, J.K., Darcy, C.J., Ford, R.E., 1990. Effects of Drought Stress and Infection by Maize-Dwarf Mosaic-Virus on Sweet Corn. Plant Dis 74(2), 147-151.

OOi, K., Ohshita, S., Ishii, I., Yahara, T., 1997. Molecular phylogeny of geminivirus infecting wild plants in Japan. Journal of Plant Research 110, 247-257.

Ostfeld, R.S., Glass, G.E., Keesing, F., 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in ecology & evolution 20(6), 328-336.

Ostfeld, R.S., Holt, R.D., 2004. Are predators good for your health? Evaluating evidence for top-down regulation of zoonotic disease reservoirs. Frontiers in Ecology and the Environment 2(1), 13-20.

Packer, C., Holt, R.D., Hudson, P.J., Lafferty, K.D., Dobson, A.P., 2003. Keeping the herds healthy and alert: implications of predator control for infectious disease. Ecol Lett 6(9), 797-802.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265(2), 218-225.

Pagan, I., Alonso-Blanco, C., Garcia-Arenal, F., 2008. Host responses in life-history traits and tolerance to virus infection in Arabidopsis thaliana. PLoS pathogens 4(8), e1000124.

Pagan, I., Gonzalez-Jara, P., Moreno-Letelier, A., Rodelo-Urrego, M., Fraile, A., Pinero, D., Garcia-Arenal, F., 2012. Effect of biodiversity changes in disease risk: exploring disease emergence in a plant-virus system. PLoS pathogens 8(7), e1002796.

Pallas, V., Garcia, J.A., 2011. How do plant viruses induce disease? Interactions and interference with host components. The Journal of general virology 92(Pt 12), 2691-2705.

Pallett, D.W., Ho, T., Cooper, I., Wang, H., 2010. Detection of Cereal yellow dwarf virus using small interfering RNAs and enhanced infection rate with Cocksfoot streak virus in wild cocksfoot grass (Dactylis glomerata). Journal of virological methods 168(1-2), 223-227.

Pantaleo, V., Szittya, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T., Burgyan, J., 2010. Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. The Plant Journal, no-no.

Pearson, H., 2008. 'Virophage' suggests viruses are alive. Nature 454(7205), 677-677.

Peng, Y., Leung, H.C., Yiu, S.M., Chin, F.Y., 2011. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27(13), i94-101.

Perefarres, F., Thierry, M., Becker, N., Lefeuvre, P., Reynaud, B., Delatte, H., Lett, J.M., 2012. Biological invasions of geminiviruses: case study of TYLCV and Bemisia tabaci in Reunion Island. Viruses 4(12), 3665-3688.

Perez-Brocal, V., Garcia-Lopez, R., Vazquez-Castellanos, J.F., Nos, P., Beltran, B., Latorre, A., Moya, A., 2013. Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. Clinical and translational gastroenterology 4, e36.

Perry, J., 1999. <Perry,1999, Red-Blue plots.pdf>.

Perry, J.N., 1995. Spatial Analysis by Distance Indices. J. Anim. Ecol. 64, 303-3014.

Perry, J.N., 1996. SADIE: software to measure and model spatial pattern for counts. Aspects Appl. Biol. 46, 95-102.

Perry, J.N., 1998. Measures of spatial pattern and spatial association for counts of insects. In: Baumgartner, J., Brandmayr, P, manly, B.F.J. (Ed.), Population and Community Ecology for Insect Management and Conservation. Balkema, Rotterdam, pp. 21-31.

Perry, J.N., Dixon, P.M., 2002. A new method to measure spatial association for ecological count data. Ecoscience 9(2), 133-141.

Perry, J.N., Winder, L., Holland, J.M., Alston, R.D., 1999. Red-blue plots for detecting clusters in count data. Ecol Lett 2(2), 106-113.

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., Abergel, C., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341(6143), 281-286.

Pielou, E.C., 1969. An introduction to mathematical ecology, viii, 286 p. pp. Wiley-Interscience, New York,.

Pignatelli, M., Moya, A., 2011. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PloS one 6(5), e19984.

Plantegenest, M., Le May, C., Fabre, F., 2007. Landscape epidemiology of plant diseases. Journal of the Royal Society, Interface / the Royal Society 4(16), 963-972.

Plumb, R.T., Thresh, J.M., Federation of British Plant Pathologists., 1983. Plant virus epidemiology : the spread and control of insect-borne viruses, vii, 377 p. pp. Blackwell Scientific Publications ;Blackwell Mosby Book Distributors, Oxford ; Boston St. Louis, Mo.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25(7), 1253-1256.

Power, A.G., 1991. Virus Spread and Vector Dynamics in Genetically Diverse Plant-Populations. Ecology 72(1), 232-241.

Power, A.G., 2008. Community ecology of plant viruses. In: Roossinck, M.J. (Ed.), Plant virus evolution. Springer

Verlag Berlin Heidelberg, pp. 15-26.

Power, A.G., Borer, E.T., Hosseini, P., Mitchell, C.E., Seabloom, E.W., 2011. The community ecology of barley/cereal yellow dwarf viruses in Western US grasslands. Virus research 159(2), 95-100.

Power, A.G., Flecker, A.S., 2003. Virus specificity in disease systems: are species redundant? In: Kareiva, P., Levin, S.A. (Ed.), The importance of specie : Perspectives on expendability and triage Princeton University Press, Princeton NJ, pp. 330-346.

Power, A.G., Mitchell, C.E., 2004. Pathogen spillover in disease epidemics. The American naturalist 164 Suppl 5, S79-89.

Prendeville, H.R., Ye, X., Jack Morris, T., Pilson, D., 2012. Virus infections in wild plant populations are both frequent and often unapparent. American journal of botany 99(6), 1033-1042.

Qi, X., Bao, F.S., Xie, Z., 2009. Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. PloS one 4(3), e4971.

Querci, M., Owens, R.A., Bartolini, I., Lazarte, V., Salazar, L.F., 1997. Evidence for heterologous encapsidation of potato spindle tuber viroid in particles of potato leafroll virus. The Journal of general virology 78 (Pt 6), 1207-1211.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306(5700), 1344-1350.

Real, L.A., 1996. Disease ecology. Ecology 77(4), 989-989.

Remold, S.K., 2002. Unapparent virus infection and host fitness in three weedy grass species. J Ecol 90(6), 967-977.

Rey, M.E.C., Ndunguru, J., Berrie, L.C., Paximadis, M., Berry, S., Cossa, N., Nuaila, V.N., Mabasa, K.G., Abraham, N., Rybicki, E.P., Martin, D., Pietersen, G., Esterhuizen, L.L., 2012. Diversity of Dicotyledenous-Infecting Geminiviruses and Their Associated DNA Molecules in Southern Africa, Including the South-West Indian Ocean Islands. Viruses 4(9), 1753-1791.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., Gordon, J.I., 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466(7304), 334-338.

Reyes, G.R., Kim, J.P., 1991. Sequence-independant, single-primer amplification (SISPA) of complex DNA populations. Mol. Cell. Probes/ 5(6), 473-481.

Rodelo-Urrego, M., Pagan, I., Gonzalez-Jara, P., Betancourt, M., Moreno-Letelier, A., Ayllon, M.A., Fraile, A., Pinero, D., Garcia-Arenal, F., 2013. Landscape heterogeneity shapes host-parasite interactions and results in apparent plant-virus codivergence. Mol Ecol 22(8), 2325-2340.

Rodriguez, R.L., Konstantinidis, K.T., 2014. Estimating coverage in metagenomic data sets and why it matters. The ISME journal.

Roossinck, M.J., 2005. Symbiosis versus competition in plant virus evolution. Nature Reviews Microbiology 3(12), 917-924.

Roossinck, M.J., 2010. Lifestyles of plant viruses. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365(1548), 1899-1905.

Roossinck, M.J., 2011a. The big unknown: plant virus biodiversity. Curr Opin Virol 1(1), 63-67.

Roossinck, M.J., 2011b. Changes in population dynamics in mutualistic versus pathogenic viruses. Viruses 3(1), 12-19.

Roossinck, M.J., 2011c. The good viruses: viral mutualistic symbioses. Nature reviews. Microbiology 9(2), 99-108.

Roossinck, M.J., 2012a. Persistent Plant Viruses: Molecular Hitchhikers or Epigenetic Elements? , 177-186.

Roossinck, M.J., 2012b. Plant Virus Metagenomics: Biodiversity and Ecology. Annual review of genetics.

Roossinck, M.J., 2013. Plant virus ecology. PLoS pathogens 9(5), e1003304.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G.A., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19, 81-88.

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Current Opinion in Virology 1(4), 289-297.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M., Varsani, A., 2012. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). The Journal of general virology 93(Pt 12), 2668-2681.

Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y., Breitbart, M., 2009. Metagenomic analysis of viruses in reclaimed water. Environmental microbiology 11(11), 2806-2820.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D.P., Breitbart, M., Varsani, A., 2013. Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. Virus research 171(1), 231-237.

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., Van Borm, S., 2014. False-Positive Results in Metagenomic Virus Discovery: A Strong Case for Follow-Up Diagnosis. Transboundary and emerging diseases.

Roux, S., 2013. Diversité, évolution, et écologie virale: des communautés aux génotypes, Université Blaise Pascal Clermont-Ferrand.

Roux S, E.F., Robin A, Ravet V, Personnic S, , Theil S, C.J., Sime-Ngando T, Debroas D, 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PloS one 7(3).

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., Enault, F., 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27(21), 3074-3075.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., Enault, F., 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. Bmc Bioinformatics 15, 76.

Ruiz-Ruiz, S., Navarro, B., Gisel, A., Pena, L., Navarro, L., Moreno, P., Di Serio, F., Flores, R., 2011. Citrus tristeza virus infection induces the accumulation of viral small RNAs (21-24-nt) mapping preferentially at the 3'-terminal region of the genomic RNA and affects the host small RNA profile. Plant molecular biology 75(6), 607-619.

Rybicki, E.P., 1990. The classification of organisms at the edge of life, or problems with virus systematics. South African Journal of Science 86, 182-186.

Rybicki, E.P., 1994. A phylogenetic and evolutionary justification for three genera of Geminiviridae. Archives of virology 139(1-2), 49-77.

Rybicki, E.P., Pietersen, G., 1999. Plant virus disease problems in the developing world. Advances in Virus Research, Vol 53 53, 127-+.

Sacristan, S., Fraile, A., Malpica, J.M., Garcia-Arenal, F., 2005. An Analysis of Host Adaptation and Its Relationship with Virulence in Cucumber mosaic virus. Phytopathology 95(7), 827-833.

Saïb, A., 2006. Les virus inertes ou vivants? Pour la science 350, 60-64.

Sanjuan, R., Moya, A., Elena, S.F., 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America 101(22), 8396-8401.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. Journal of virology 84(19), 9733-9748.

Sastry, K.S., 2013. Seed-borne plant virus diseases.

Saunders, K., Bedford, I.D., Yahara, T., Stanley, J., 2003. The earliest recorded plant virus disease. Nature 422(6934), 831-831.

Saunders, K., Stanley, J., 1999. A nanovirus-like DNA component associated with yellow vein disease of Ageratum conyzoides: evidence for interfamilial recombination between plant DNA viruses. Virology 264(1), 142-152.

Schmidt, T.M., DeLong, E.F., Pace, N.R., 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. Journal of bacteriology 173(14), 4371-4378.

Schmieder, R., Edwards, R., 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS one 6(3), e17288.

Scholz, M.B., Lo, C.C., Chain, P.S., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current opinion in biotechnology 23(1), 9-15.

Schrotenboer, A.C., Allen, M.S., Malmstrom, C.M., 2011. Modification of native grasses for biofuel production may increase virus susceptibility. Gcb Bioenergy 3(5), 360-374.

Seabloom, E.W., Harpole, W.S., Reichman, O.J., Tilman, D., 2003. Invasion, competitive dominance, and resource use by exotic and native California grassland species. Proceedings of the National Academy of Sciences of the United States of America 100(23), 13384-13389.

Shannon, C.E., 1948. A mathematical theory of comunication. The Bell System Technical Journal 27, 379-423.

Sharpton, T.J., 2014. An introduction to the analysis of shotgun metagenomic data. Frontiers in plant science 5, 209.

Shepherd, D.N., Martin, D.P., Lefeuvre, P., Monjane, A.L., Owor, B.E., Rybicki, E.P., Varsani, A., 2008. A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. Journal of virological methods 149(1), 97-102.

Sibley, C.D., Peirano, G., Church, D.L., 2012. Molecular methods for pathogen and microbial community detection and characterization: current and potential application in diagnostic microbiology. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 12(3), 505-521.

Sikorski, A., Massaro, M., Kraberger, S., Young, L.M., Smalley, D., Martin, D.P., Varsani, A., 2013. Novel myco-like DNA viruses discovered in the faecal matter of various animals. Virus research 177(2), 209-216.

Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688.

Singh, K.M., Ahir, V.B., Tripathi, A.K., Ramani, U.V., Sajnani, M., Koringa, P.G., Jakhesara, S., Pandya, P.R., Rank, D.N., Murty, D.S., Kothari, R.K., Joshi, C.G., 2012. Metagenomic analysis of Surti buffalo (Bubalus bubalis) rumen: a preliminary study. Molecular biology reports 39(4), 4841-4848.

Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., Allaby, R.G., 2014. A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. Scientific reports 4, 4003.

Solonenko, S.A., Ignacio-Espinoza, J.C., Alberti, A., Cruaud, C., Hallam, S., Konstantinidis, K., Tyson, G., Wincker, P., Sullivan, M.B., 2013. Sequencing platform and library preparation choices impact viral metagenomes. Bmc Genomics 14, 320.

St Clair, D.A., 2010. Quantitative disease resistance and quantitative resistance Loci in breeding. Annual review of phytopathology 48, 247-268.

Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., Bergelson, J., 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 400(6745), 667-671.

Stanley, J., Markham, P.G., Callis, R.J., Pinner, M.S., 1986. The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. The EMBO journal 5(8), 1761-1767.

Stanley, W.M., 1935. Isolation of a Crystalline Protein Possessing the Properties of Tobacco-Mosaic Virus. Science 81(2113), 644-645.

Stapleton, A.E., 2014. A biologist,a statistician,and a bioinformatician walk into a conference room...and walk out with a great metagenomics project plan. Frontiers in plant science.

Steddom, K., Jones, D., Rudd, J., Rush, C., 2005. Analysis of field plot images with segmentation analysis, effect of glare and shadows. Phytopathology 95(6), S99-S99.

Steinmann, V.W., Porter, J.M., 2002. Phylogenetic relationships in Euphorbieae (Euphorbiaceae) based on its and ndhF sequence data. Ann Mo Bot Gard 89(4), 453-490.

Stern, N., Taylor, C., 2007. Economics. Climate change: risk, ethics, and the Stern Review. Science 317(5835), 203-204.

Stubbs, L.L., Grogan, R.G., 1963. Necrotic yellows: A newly recognized virus disease of lettuce. Australian journal of agricultural reasearch 14(4), 439-459.

Sutherst, R.W., Constable, F., Finlay, K.J., Harrington, R., Luck, J., Zalucki, M.P., 2011. Adapting to crop pest and pathogen risks under a changing climate. Wires Clim Change 2(2), 220-237.

Syller, J., Marczewski, W., 2001. Potato leafroll virus-assisted aphid transmission of potato spindle tuber viroid to potato leafroll virus-resistant potato. Journal of Phytopathology-Phytopathologische Zeitschrift 149(3-4), 195-201.

Szathmary, E., 1992. Viral sex, levels of selection, and the origin of life. Journal of theoretical biology 159(1), 99-109.

Szittya, G., Moxon, S., Pantaleo, V., Toth, G., Rusholme Pilcher, R.L., Moulton, V., Burgyan, J., Dalmay, T., 2010. Structural and functional analysis of viral siRNAs. PLoS pathogens 6(4), e1000838.

Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant molecular biology 17(5), 1105-1109.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28(10), 2731-2739.

Thackray, D.J., Smith, L.J., Cheng, Y., Perry, J.N., Jones, R.A.C., 2002. Effect of strain-specific hypersensitive resistance on spatial patterns of virus spread. Ann Appl Biol 141(1), 45-59.

Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E., Thompson, F.L., 2013. Microbial genomic taxonomy. Bmc Genomics 14, 913.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic acids research 22(22), 4673-4680.

Thresh, J.M., 2006. Crop viruses and virus diseases: A global perspective. Nato Sci Peace Secur, 9-32.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., Rohwer, F., 2009. Laboratory procedures to generate viral metagenomes. Nature protocols 4(4), 470-483.

Tse, H., Tsang, A.K., Tsoi, H.W., Leung, A.S., Ho, C.C., Lau, S.K., Woo, P.C., Yuen, K.Y., 2012. Identification of a novel bat papillomavirus by metagenomics. PloS one 7(8), e43986.

Turner, M.G., 2005. Landscape ecology: What is the state of science? Annu. Rev. Ecol. Evol. Syst. 36, 319-344.

Uma, B., Rani, T.S., Podile, A.R., 2011. Warriors at the gate that never sleep: Non-host resistance in plants. J Plant Physiol 168(18), 2141-2152.

van den Brand, J.M., van Leeuwen, M., Schapendonk, C.M., Simon, J.H., Haagmans, B.L., Osterhaus, A.D., Smits, S.L., 2012. Metagenomic analysis of the viral flora of pine marten and European badger feces. Journal of virology 86(4), 2360-2365.

Van Regenmortel, M.H., 1989. Applying the species concept to plant viruses. Archives of virology 104(1-2), 1-17.

Van Regenmortel, M.H., 2000. Introduction to the species concept in virus taxonomy. In: Van Regenmortel, M.H., Fauquet, C.M., Bishop, D.H.e.*a.* (Eds.), Seventh Report of the ICTV. Academic Press, San Diego, pp. 3-16.

Van Regenmortel, M.H., Ackermann, H.W., Calisher, C.H., Dietzgen, R.G., Horzinek, M.C., Keil, G.M., Mahy, B.W., Martelli, G.P., Murphy, F.A., Pringle, C., Rima, B.K., Skern, T., Vetten, H.J., Weaver, S.C., 2013. Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. Archives of virology 158(5), 1115-1119.

Van Regenmortel, M.H., Bishop, D.H., Fauquet, C.M., Mayo, M.A., Maniloff, J., Calisher, C.H., 1997. Guidelines to the demarcation of virus species. Archives of virology 142(7), 1505-1518.

Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Murilo Zerbini, F., Brown, J.K., 2014a. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Archives of virology 159(7), 1873-1882.

Varsani, A., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Brown, J.K., Murilo Zerbini, F., Martin, D.P., 2014b. Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Archives of virology.

Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P., Martin, D.P., 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. Virology journal 6, 36.

Vazquez-Castellanos, J.F., Garcia-Lopez, R., Perez-Brocal, V., Pignatelli, M., Moya, A., 2014. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. Bmc Genomics 15, 37.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667), 66-74.

Victoria, J.G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S., Delwart, E., 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. Journal of virology 83(9), 4642-4651.

Villarreal, L.P., 2008. Are viruses alive?, . Vol. 2014.

Vincent, S.J., Coutts, B.A., Jones, R.A., 2014. Effects of introduced and indigenous viruses on native plants: exploring their disease causing potential at the agro-ecological interface. PloS one 9(3), e91224.

Vuillaume, F., Thebaud, G., Urbino, C., Forfert, N., Granier, M., Froissart, R., Blanc, S., Peterschmitt, M., 2011. Distribution of the phenotypic effects of random homologous recombination between two virus species. PLoS pathogens 7(5), e1002028.

Walker, A., 2010. Gut metagenomics goes viral. Nature reviews. Microbiology 8(12), 841.

Ward, D.M., Cohan, F.M., Bhaya, D., Heidelberg, J.F., Kuhl, M., Grossman, A., 2008. Genomics, environmental genomics and the issue of microbial species. Heredity 100(2), 207-219.

Webster, C.G., Coutts, B.A., Jones, R.A.C., Jones, M.G.K., Wylie, S.J., 2007. Virus impact at the interface of an ancient ecosystem and a recent agroecosystem: studies on three legume-infecting potyviruses in the southwest Australian floristic region. Plant Pathol 56(5), 729-742.

Wesche, P.L., Gaffney, D.J., Keightley, P.D., 2004. DNA sequence error rates in Genbank records estimated using the mouse genome as a reference. DNA sequence : the journal of DNA sequencing and mapping 15(5-6), 362-364.

Whittaker, R.H., 1960. Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs 30, 279-338.

Williamson, G., 2011. Euphorbia caput-medusae L.-a journey from the foggy Cape of Storms to the arid wind-blasted sands of the Namib Desert. Euphorbia World 7(2).

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., Rohwer, F., 2009a. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PloS one 4(10), e7370.

Willner, D., Thurber, R.V., Rohwer, F., 2009b. Metagenomic signatures of 86 microbial and viral metagenomes. Environmental microbiology 11(7), 1752-1766.

Winder, L., Alexander, C.J., Holland, J.M., Woolley, C., Perry, J.N., 2001. Modelling the dynamic spatio-temporal response of predators to transient prey patches in the field. Ecol Lett 4(6), 568-576.

Wisler, G.C., Norris, R.E., 2005. Interactions between weeds and cultivated plants as related to management of plant pathogens. Weed Sci 53(6), 914-917.

Wolda, H., 1981. Similarity indices, sample size and diversity. Oecologia 50, 296-302.

Wommack, K.E., Bhavsar, J., Ravel, J., 2008. Metagenomics: Read length matters. Applied and environmental microbiology 74(5), 1453-1463.

Wong, K., Fong, T.T., Bibby, K., Molina, M., 2012. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environ Int 45, 151-164.

Wooley, J.C., Ye, Y.Z., 2010. Metagenomics: Facts and Artifacts, and Computational Challenges. J Comput Sci Tech-Ch 25(1), 71-81.

Woolhouse, M.E., Webster, J.P., Domingo, E., Charlesworth, B., Levin, B.R., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nature genetics 32(4), 569-577.

Wren, J.D., Roossinck, M.J., Nelson, R.S., Scheets, K., Palmer, M.W., Melcher, U., 2006. Plant virus biodiversity and ecology. PLoS biology 4(3), 314-315.

Wright, C.F., Morelli, M.J., Thebaud, G., Knowles, N.J., Herzyk, P., Paton, D.J., Haydon, D.T., King, D.P., 2011. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. Journal of virology 85(5), 2266-2275.

Wu, Q.F., Luo, Y.J., Lu, R., Lau, N., Lai, E.C., Li, W.X., Ding, S.W., 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proceedings of the National Academy of Sciences of the United States of America 107(4), 1606-1611.

Wyant, P.S., Stephan Strohmeier , Benjamin Schäfer , Björn Krenz , Iraildes Pereira Assunção , Gaus Silvestre de Andrade Lima , Holger Jeske, 2012. Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. Virology.

Wylie, S.J., Li, H., Dixon, K.W., Richards, H., Jones, M.G., 2013. Exotic and indigenous viruses infect wild populations and captive collections of temperate terrestrial orchids (Diuris species) in Australia. Virus research 171(1), 22-32.

Wylie, S.J., Luo, H., Li, H., Jones, M.G., 2011. Multiple polyadenylated RNA viruses detected in pooled cultivated and wild plant samples. Archives of virology.

Xu, P., Chen, F., Mannas, J.P., Feldman, T., Sumner, L.W., Roossinck, M.J., 2008. Virus infection improves drought tolerance. New Phytologist 180(4), 911-921.

Xu, Y., Huang, L., Fu, S., Wu, J., Zhou, X., 2012. Population diversity of rice stripe virus-derived siRNAs in three different hosts and RNAi-based antiviral immunity in Laodelphgax striatellus. PloS one 7(9), e46238.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J., Wang, J., Jin, Q., 2011. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. Journal of clinical microbiology 49(10), 3463-3469.

Yilmaz, S., Allgaier, M., Hugenholtz, P., 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature methods 7(12), 943-944.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C.S., Li, H., Mashiyama, S.T., Joachimiak, M.P., van Belle, C., Chandonia, J.M., Soergel, D.A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B.J., Bafna, V., Friedman, R., Brenner, S.E., Godzik, A., Eisenberg, D., Dixon, J.E., Taylor, S.S., Strausberg, R.L., Frazier, M., Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS biology 5(3), e16.

Yu, X., Li, B., Fu, Y.P., Jiang, D.H., Ghabrial, S.A., Li, G.Q., Peng, Y.L., Xie, J.T., Cheng, J.S., Huang, J.B., Yi, X.H., 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proceedings of the National Academy of Sciences of the United States of America 107(18), 8387-8392.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.Q., Wei, C.L., Soh, S.W., Hibberd, M.L., Liu, E.T., Rohwer, F., Ruan, Y., 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS biology 4(1), e3.

Abstract

Ecology, diversity, and discovery of plant viruses over space and time across two agro-ecosystems using geo-metagenomics

Our knowledge about plant virus diversity in nature is still limited. Indeed, studies of plant-virus interactions have primarily focused on cultivated areas. This lack of knowledge about patterns of virus diversity and distribution in nature is hampering our understanding of plant virus ecology and evolution in the long term. In addition, this scarcity of knowledge does not allow to fully understand, model and predict the micro- and/or macro-evolutionary processes that are taking place across the agro-ecosystem. Consequently, it is still difficult to quantify the impact of human activities (agricultural intensification, plants transport, climate change, etc.) on host-pathogen interactions.

We developed a new metagenomics approach, the so-called geo-metagenomics approach, in order to provide information about the virus biodiversity, the prevalence of unknown and asymptomatic viruses and the spatial distributions of those plant viruses in two pilot ecosystems: the Western Cape Region of South Africa and the Camargue region in France. This approach provides geographically tagged cDNA from known and unknown viruses, and further allows linking viral sequences obtained by the metagenomic approach to a specific host, and hence to geographic coordinates. The objectives of this study were to assess (i) if wild areas can be considered as reservoir of plant virus biodiversity (ii) if there exists patterns of spatio-temporal distribution of plant viruses at the agro-ecosystem scale and (ii) if ecological parameters can account for these distributions.

This new approach allowed us to estimate plant virus diversity associated with two agro-ecosystems. Patterns of spatial and temporal distribution of several viral families have been highlighted. Plant virus prevalences associated with cultivated areas were found to be significantly greater than those associated with non-cultivated areas for three out of the four sampling surveys. Furthermore, exotic plants from South African fynbos showed significantly higher prevalence than native plants. These results emphasize the direct or indirect impact of human activity on plant virus dynamics at the agroecosystem scale.

This study also led to the discovery of hundreds of potential new viral species, including three new species belonging to the family *Geminiviridae*. These species belong to a new genus of the *Geminiviridae* family that we tentatively named Capulavirus. This discovery sheds a new light on the evolutionary history of geminiviruses (recombination, genomic features of their common ancestor). This new genus contains four species, including two species isolated from wild plants (Euphorbia caput-medusae latent virus and *Plantago Capulavirus*) and two species recovered from crops (Alfalfa leaf curl virus and French bean severe leaf curl virus). Finally, our results suggest that aphids may transmit virus species from this new genus, which has never been described so far for geminiviruses vection. The potential emergence of this new genus is finally discussed.

Keywords: Metagenomics, agro-ecosystem, plant viruses, diversity, ecology, evolution, recombination, *Geminiviridae*, Capulavirus, exotic/native plants, fynbos, Camargue.

Résumé

Ecologie, diversité, et découverte de phytovirus à l'échelle de deux agro-écosystèmes dans un cadre spatio-temporel à l'aide de la géo-métagénomique

La connaissance de la diversité des phytovirus en milieu sauvage reste limitée. Les études concernant les interactions plantes-virus se sont en effet principalement focalisées sur les milieux cultivés. Ce manque de connaissance des milieux naturels et des interactions plante-virus qui s'y déroulent représente un écueil dans notre compréhension de l'écologie et de l'évolution des phytovirus sur le long terme. Cette quasi-absence de connaissance ne permet en outre pas de totalement comprendre, modéliser et prédire les processus micro- et/ou macro-évolutifs qui se mettent en place à l'échelle de l'agro-écosystème. Il est notamment encore difficile de quantifier l'impact des activités humaines (intensification de l'agriculture, transport de plantes, changement du climat, etc.) sur les interactions hôtes-agents pathogènes.

Une approche de géo-métagénomique a été développée dans deux agro-écosystèmes (le fynbos en Afrique du Sud et la Camargue) sur un pas de temps de deux ans. Cette approche nous a permis de ré-attribuer chaque séquence virale à son hôte géolocalisé. L'objectif de ce travail était d'évaluer (i) si le milieu sauvage constitue un réservoir de biodiversité phytovirale (ii) si il existe des patrons de distribution spatio-temporelle des phytovirus dans l'agro-écosystème et (ii) si des paramètres écologiques permettent d'expliquer ces distributions.

Grâce à cette nouvelle approche, une estimation de la diversité phytovirale associée aux deux agro-écosystèmes a pu être obtenue. Des patrons de distribution spatio-temporelle de plusieurs familles virales ont pu être mis en évidence. Les prévalences phytovirales associées au milieu cultivé se sont avérées être significativement plus importantes que celles associées au milieu non-cultivé pour trois des quatre campagnes d'échantillonnage. Par ailleurs, les plantes exotiques du fynbos sud-africain ont présenté des prévalences phytovirales significativement plus élevées que celles des plantes indigènes. Ces résultats soulignent l'impact direct ou indirect de l'activité humaine sur les dynamiques phytovirales à l'échelle de l'agro-écosystème.

Cette étude a également mené à la découverte potentielle de centaines de nouvelles espèces virales, dont trois nouvelles espèces appartenant à la famille des *Geminiviridae*. Ces espèces appartiennent à un nouveau genre des *Geminiviridae* que nous avons nommé Capulavirus. Cette découverte nous a permis de mieux estimer certains paramètres liés à l'histoire évolutive des géminivirus (recombinaison, caractéristiques de leur ancêtre commun). Ce nouveau genre contient quatre espèces dont deux issues de plantes sauvages (Euphorbia caput-medusae latent virus et *Plantago Capulavirus*) et deux de plantes cultivées (Alfafa leaf curl virus et French bean severe leaf curl virus). Par ailleurs, nous avons obtenu des résultats préliminaires suggérant la transmission de ce nouveau genre par puceron, insecte qui n'a jamais été décrit comme vecteur de géminivirus. Ces découvertes nous ont amené à émettre des hypothèses sur l'émergence potentielle de ce nouveau genre à l'échelle mondiale.

Mots clés : Métagénomique, agro-écosystème, phytovirus, diversité, écologie, évolution, recombinaison, *Geminiviridae*, Capulavirus, plantes exotiques, plantes indigènes, fynbos, Afrique du Sud, Camargue.